JEOVANI SCHMITT

PRÉ-PROCESSAMENTO PARA A MINERAÇÃO DE DADOS: USO DA ANÁLISE DE COMPONENTES PRINCIPAIS COM ESCALONAMENTO ÓTIMO

FLORIANÓPOLIS – SC
2005

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Jeovani Schmitt

PRÉ-PROCESSAMENTO PARA A MINERAÇÃO DE DADOS: USO DA ANÁLISE DE COMPONENTES PRINCIPAIS COM ESCALONAMENTO ÓTIMO

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Dr. Dalton Francisco de Andrade

Co-Orientador: Prof. Dr. Pedro Alberto Barbetta

Florianópolis, Outubro de 2005.

PRÉ-PROCESSAMENTO PARA A MINERAÇÃO DE DADOS: USO DA ANÁLISE DE COMPONENTES PRINCIPAIS COM ESCALONAMENTO ÓTIMO

Jeovani Schmitt

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, Área de Concentração Sistemas de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Dr. Raul Sidnei Wazlawick

Coordenador do Curso

_____

Prof. Dr. Dalton Francisco de Andrade (Orientador)

_____

Prof. Dr. Pedro Alberto Barbetta (Co-Orientador)

Banca Examinadora

_____

Prof. Dr. Adriano Ferreti Borgatto - UFSC

_____

Prof. Dra. Lúcia Pereira Barroso – IME - USP

_____

Prof. Dr. Paulo José Oglirari - UFSC

“Há muito que saber, e é pouco o viver,
e não se vive se não se aprende.”
(José Antônio Marina)

Aos meus pais, José e Celita, pelos ensinamentos de vida transmitidos.

A vocês, minha eterna gratidão e reconhecimento.

AGRADECIMENTOS

A Deus.

A Universidade Federal de Santa Catarina, que através do Programa de Pós Graduação em Ciência da Computação, viabilizou o curso.

Ao meu orientador, Professor Dr. Pedro Alberto Barbetta, pelo interesse que sempre demonstrou. Foi extremamente importante sua participação para o êxito deste trabalho. Muito obrigado também pelos conhecimentos transmitidos.

À minha esposa, Kátia, pelo incentivo dado ao longo da realização do trabalho.

Ao Professor Dr. Dalton Francisco de Andrade, pela presteza aos assuntos relacionados a este trabalho.

Aos Professores da Banca Examinadora, Dra. Lúcia Pereira Barroso, Dr. Adriano Borgatto e Dr. Paulo José Ogliari, pelas orientações e correções sugeridas para a finalização do trabalho.

Aos amigos que conheci durante o curso, principalmente Jacqueline Uber, sendo sempre muito companheira.

A todos que contribuíram direta ou indiretamente para a realização deste trabalho.

RESUMO

A mineração de dados em grandes bases pode requerer alto tempo computacional. Além do mais, é comum as bases de dados conterem variáveis mensuradas em diferentes níveis: intervalar, ordinal e nominal. Neste caso, técnicas desenvolvidas para variáveis quantitativas não poderiam ser aplicadas sob as variáveis originais. Como exemplo, pode-se citar a análise de agrupamentos pelo método das k-médias. Este exige que as variáveis de entradas sejam quantitativas.

Este trabalho apresenta uma metodologia para a fase do pré-processamento em mineração de dados, que utiliza a análise de componentes principais (ACP) com escalonamento ótimo (EO). O pré-processamento é uma etapa fundamental que pode melhorar a performance dos algoritmos de análise, através da redução de dimensionalidade. O escalonamento ótimo permite analisar bases que contenham variáveis observadas em diferentes níveis de mensuração.

Através da ACP é possível obter uma redução das variáveis originais em um número de componentes principais, gerando novas coordenadas, menor que o número de variáveis originais. As novas coordenadas podem ser utilizadas na mineração de dados propriamente dita, em tarefas como agrupamentos, classificação entre outras. Essas tarefas podem ser realizadas por métodos estatísticos ou computacionais, como redes neurais, algoritmos genéticos entre outros.

A metodologia proposta foi testada em uma base de dados com 118.776 registros de pessoas, pesquisadas pelo Instituto Brasileiro de Geografia e Estatística - IBGE, contendo 13 variáveis observadas em diferentes níveis de mensuração. Através da ACP com EO, as 13 variáveis foram reduzidas a 6 componentes principais, preservando ainda 77% da variabilidade original. Sob o novo conjunto de coordenadas foi aplicada a análise de agrupamentos, utilizando o algoritmo das k-médias para a separação dos grupos, com o objetivo de ilustrar uma tarefa comum em mineração de dados, a identificação de grupos, onde foi possível descrever 6 subgrupos ou *clusters*.

Palavras-Chave: Componentes Principais, Escalonamento Ótimo, Mineração de Dados, Pré-Processamento.

ABSTRACT

Data mining in large databases might require high computational time. Besides, in databases are common to have variables measured in different levels: interval, ordinal and nominal. In this case, techniques developed for quantitative variables could not be applied to the original variables. As an example, the clusters analysis may be mentioned by k-means method. This one demands the variables of entrances are quantitative.

This work presents a methodology for preprocessing phase which uses the principal components analysis (PCA) with optimal scaling (OS). The preprocessing is a fundamental stage and it might be used to improve the performance analysis algorithms, reducing the dimensionality. The optimal scaling allow to analyze databases that contain variables observed in different measurement levels.

Through PCA, a reduction of the original variables is possible to be obtained in a number of main components, generating new coordinates, smaller than the number of originals variables. The new coordinates can be used in data mining, in tasks as clusters, classification and others. Those tasks may be accomplished by statistical or computationals methods, as neural network, genetic algorithms and others.

The proposed methodology was tested in a database with 118.776 registers of people, researched by Instituto Brasileiro de Geografia e Estatística - IBGE, containing 13 variables observed in different mensurement levels. Through PCA with OS, the 13 variables were reduced to 6 main components, still preserving 77% of the original variability. Under the new group of coordinates the analysis of groupings was applied, using the k-means algorithm for the separation of the groups, which aims is to illustrate a common task in data mining, identify groups, where 6 groups or clusters were possible to be described.

Key-words: Principal Components, Optimal Scaling, Data Mining, Preprocessing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

$\mathbf{B}$	Matriz de ordem n.m contendo os valores ordenados da matriz $\mathbf{Z}^{*}_{nxm}$
$\mathbf{D}_k$	Matriz diagonal relativa aos k componentes principais
$\mathbf{D}_k^{-1}$	Matriz inversa da matriz diagonal relativa aos k componentes principais
$\mathbf{F}$	Matriz dos escores fatoriais (observações) – *factor scores*
$\mathbf{G}$	Matriz indicadora binária para o escalonamento ótimo
$\mathbf{L}$	Matriz das cargas fatoriais (variáveis) – *factor loadings*
$\mathbf{I}$	Matriz Identidade
k	número de variáveis ou dimensões das variáveis transformadas
m	número de variáveis, colunas ou dimensões
n	número de observações, casos ou linhas
n´	tamanho da amostra
$\mathbf{O}_m$	Matriz nula
$\mathbf{R}$	Matriz de correlações amostrais
$\mathbb{R}^m$	espaço de representação de m dimensões
$\mathbb{R}^k$	subespaço de representação de k dimensões
s_j	desvio padrão da coluna j
$\mathbf{S}$	Matriz de covariâncias
tr	traço da matriz
$\mathbf{u}$	autovetor não normalizado
$\mathbf{v}_1$	vetor que representa a primeira direção de máxima variabilidade
$\mathbf{v}_2$	vetor que representa a segunda direção de máxima variabilidade
$\mathbf{v}_1^t$	vetor transposto de $\mathbf{v}_1$
$\mathbf{v}_2^t$	vetor transposto de $\mathbf{v}_2$
$\mathbf{v}_m$	vetor que representa a m-ésima direção
$\mathbf{v}_j$	Autovetor j normalizado associado ao autovalor λ_j
$\mathbf{v}_j^t$	Transposta do Autovetor j normalizado associado ao autovalor λ_j

$\mathbf{V}$	Matriz dos autovetores normalizados
$\bar{x}_j$	média da coluna j
$\mathbf{X}$	Matriz dos dados observados
x_{ij}	elemento da matriz **X** que representa a observação i na variável j
$\mathbf{x}_i$	Representação do vetor ponto-linha
$\hat{\mathbf{x}}_i$	projeção perpendicular de $\mathbf{x}_i$ sobre o subespaço vetorial
y_{1i}, y_{2i}	coordenadas do ponto projetado $\hat{\mathbf{x}}_i$ no subespaço vetorial
$\mathbf{Y}$	Matriz das novas coordenadas determinadas pelas componentes principais
$\mathbf{Z}$	Matriz dos dados padronizados
$\mathbf{Z}^t$	Transposta da Matriz dos dados padronizados
$\mathbf{Z}^{"}$	Matriz estimada de escalonamento ótimo
$\mathbf{Z}^*$	Matriz de dados normalizados
$\hat{\mathbf{Z}}$	Matriz de ordem n.m contendo os valores ordenados da matriz $\mathbf{Z}^{"}_{nxm}$
$\mathbf{Z}^G$	Matriz dos valores não normalizados do escalonamento obtido da operador de mínimos quadrados sob a matriz $\hat{\mathbf{Z}}$
λ_j	Autovalor j

SUMÁRIO

1. INTRODUÇÃO

Contextualização do problema

Desde o primeiro computador, o Eniac em 1940, que pesava toneladas, ocupava um andar inteiro de um prédio e tinha uma programação complicada, muito se evoluiu na questão de tamanho dos computadores, linguagens de programação, capacidade de armazenar e transmitir dados.

De alguma maneira a era digital facilita o acúmulo de dados. Por outro lado, os dados por si só não revelam conhecimentos para pesquisadores, administradores e para as pessoas que lidam com informação. A necessidade de organizar os dados para que seja possível transformá-los em informações úteis, direciona para o que atualmente se conhece como mineração de dados. Muitas vezes, as informações são extraídas dos próprios dados contidos em grandes bases, com muitas variáveis e registros.

Segundo FERNANDEZ (2003,p.1), a mineração de dados automatiza o processo de descobrir relações e padrões nos dados e apresenta resultados que poderão ser utilizados em um sistema de suporte de decisão automatizado, ou acessado por tomadores de decisão.

As técnicas de mineração de dados têm suas raízes na inteligência artificial (IA) e na estatística, e vêm contribuir para a descoberta de regras ou padrões, prever futuras tendências e comportamentos ou conhecer grupos similares. Portanto, proporcionam às empresas e pesquisadores uma ferramenta útil na tomada de decisão.

O principal foco da presente pesquisa está nas situações de mineração de dados em que se têm muitas variáveis mensuradas em diferentes níveis, e que se planeja aplicar técnicas quantitativas, podendo exigir muito tempo computacional.

1.2 Objetivos

1.2.1 Objetivo geral

Apresentar uma metodologia para a fase do pré-processamento em mineração de dados, agregando as técnicas de escalonamento ótimo e componentes principais, permitindo que as variáveis de entrada sejam observadas em diferentes níveis de mensuração (ordinal, nominal ou intervalar), com a finalidade de obter uma redução na dimensionalidade do conjunto de dados e no tempo computacional.

1.2.2 Objetivos específicos

- Mostrar a importância da redução da dimensionalidade em mineração de dados;
- Descrever um algoritmo que realiza a análise de componentes principais com escalonamento ótimo;
- Apresentar uma metodologia para pré-processamento em mineração de dados que inclui técnicas de amostragem, componentes principais (ACP) e escalonamento ótimo (EO).
- Aplicar a metodologia proposta numa base de dados do Censo 2000 do Instituto Brasileiro de Geografia e Estatística – IBGE.

1.3 Métodos de desenvolvimento da pesquisa

No desenvolvimento da pesquisa será utilizado o método analítico, com base na revisão de literatura, agregando, principalmente, técnicas estatísticas, mas num enfoque de pré-processamento para a mineração de dados.

A metodologia proposta será aplicada numa base de dados extraída do Censo 2000 do IBGE, com 118.776 registros e 13 variáveis, mensuradas em diferentes níveis (ordinal, nominal e intervalar).

1.4 Justificativa

A análise de grandes bases de dados, realizada com a finalidade de extrair informações úteis, gerar conhecimento para tomadas de decisões, e é a essência da mineração de dados. Mas para realizar esta análise de maneira eficiente, a base de dados deve passar por uma etapa de pré-processamento (HAN e KAMBER, 2001, p.108).

Num aspecto mais amplo, o pré-processamento compreende um estudo detalhado da base de dados, identificando tipos de variáveis, tamanho da base, qualidade dos dados (presença de valores faltantes e discrepantes), transformações, padronização e métodos que possam contribuir para tornar a mineração mais eficiente.

Dependendo da tarefa de mineração a ser realizada (sumarização, agrupamento, associação, classificação ou predição), determinada técnica é mais apropriada. Fatores como a escolha da técnica, o tamanho da base de dados (número de variáveis e registros), o algoritmo utilizado pela técnica (se é iterativo ou não), influenciarão no tempo computacional.

Segundo MANLY (2005, p.130), alguns algoritmos de análise de agrupamentos iniciam através de uma análise de componentes principais (ACP), reduzindo as variáveis originais num número menor de componentes, reduzindo o tempo computacional.

A ACP, como etapa de pré-processamento para outras tarefas, é indicada como uma técnica para melhorar o desempenho do algoritmo de retropropagação em Redes Neurais (RN). O algoritmo de retropropagação ou *backpropagation* é um algoritmo de aprendizagem muito utilizado em RN. Segundo HAYKIN (1999, p.208), este algoritmo tem sua performance melhorada se as variáveis não forem correlacionadas. Esta tarefa de gerar variáveis não correlacionadas pode ser feita pela ACP.

Outros aspectos importantes da mineração de dados são em relação à mensuração das variáveis, a escolha da técnica e do algoritmo a ser utilizado numa tarefa de mineração.

Considerando a técnica de agrupamentos, PRASS (2004) apresenta um estudo comparativo entre vários algoritmos disponíveis na Análise de Agrupamentos. Entre eles, um dos algoritmos mais conhecidos e utilizados na análise de agrupamentos, o k-médias, exige que todas as variáveis de análise sejam numéricas ou binárias (BERRY e

LINOFF, 2004, p.359). Uma solução para este problema, apresentada por PRASS (2004, p. 28-32), é o uso de transformações nas variáveis, através da criação de variáveis indicadoras, conforme o nível de mensuração da variável. Outra solução, que será utilizada neste trabalho, é utilizar o escalonamento ótimo, que permite gerar novas variáveis quantitativas, a partir de variáveis com diferentes níveis de mensuração (MEULMANN, 2000 p.2.).

A metodologia de pré-processamento para a mineração de dados, proposta neste trabalho, inclui a utilização da análise de componentes principais com escalonamento ótimo, e se justifica pelas seguintes razões:

- permite ter como entrada bases de dados com variáveis mensuradas em diferentes níveis, gerando variáveis quantitativas como saída, não correlacionadas, possibilitando aplicar técnicas para variáveis quantitativas na fase da mineração de dados.

- permite reduzir o número de variáveis na fase da mineração de dados, contribuindo assim para a redução do tempo computacional de processamento.

1.5 Limitação da pesquisa

A metodologia a ser proposta poderá ser usada para diferentes objetivos finais (tarefas de mineração de dados), sendo mais apropriada para soluções que necessitam de métodos de aprendizagem não-supervisionados, tais como problemas de sumarização, associação e agrupamentos.

Para problemas de aprendizagem supervisionada, tais como, classificação, predição ou previsão, a metodologia aqui proposta pode precisar de ajustes, que não serão discutidos neste trabalho.

Também não serão estudadas, para cada tarefa específica da mineração de dados, as vantagens e desvantagens do uso da ACP com EO como etapa preliminar de análise, assim como o tratamento de valores faltantes (*missing values*), limitando-se a excluir o registro.

1.6 Estrutura da dissertação

A dissertação está estruturada em 7 capítulos, onde são apresentados os conceitos e técnicas aplicadas neste trabalho.

O primeiro capítulo apresenta uma breve introdução, contextualização do problema, objetivos, metodologia , justificativa e a limitação da pesquisa.

O segundo, terceiro e quarto capítulos contêm uma revisão de literatura abordando, respectivamente, alguns conceitos em mineração de dados, análise de componentes principais e a análise de componentes principais com escalonamento ótimo.

No quinto capítulo é apresentada a metodologia proposta de pré-processamento em mineração de dados.

O sexto capítulo contém uma aplicação da metodologia proposta, que utiliza uma grande base de dados de registros de pessoas do Censo 2000, pesquisada pelo Instituto Brasileiro de Geografia e Estatística – IBGE.

Finalmente, o sétimo capítulo traz as conclusões sobre o trabalho e sugestões para pesquisas futuras.

Capítulo 2

2. MINERAÇÃO DE DADOS

Neste capítulo serão abordadas algumas considerações importantes em mineração de dados, necessárias para o desenvolvimento do trabalho.

2.1 Surgimento da mineração de dados

A evolução da economia para a globalização, possibilitando explorar novos mercados, faz crescer cada vez mais a necessidade de informação para o gerenciamento de negócios, controle da produção, análise de mercado e tomada de decisões. Se por um lado a economia cresce, as empresas vendem mais, fazem mais negócios, conquistam novos mercados, a quantidade de dados gerados por estas transações também aumenta. O gerenciamento e o controle do negócio começa a ficar mais complexo em virtude desta grande quantidade de informações.

Surge na década de 80 uma arquitetura chamada *Data Warehouse* (DW), que é capaz de armazenar diferentes bases de dados. Segundo HAN e KAMBER (2001, p.12), *data warehouse* é um repositório de dados coletados de diferentes fontes, armazenados sob um esquema unificado. Os DW são construídos por um processo de limpeza dos dados, transformação, integração, cargas e atualização periódica, com objetivo de organizar da melhor forma a base de dados.

Para extrair informação destes repositórios surge a mineração de dados, que pode ser entendida como a exploração e análise de grandes quantidades de dados, de maneira automática ou semi-automática, com o objetivo de descobrir padrões e regras relevantes (BERRY e LINOFF, 2004, p.7).

A extração de informações das bases de dados pode ser feita através dos *softwares* de mineração de dados, como por exemplo, Bussines Objects, SPSS e SAS.

Para HAN e KAMBER (2001,p.5;7), a mineração de dados está ligada a descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases* – KDD). A

mineração de dados é uma etapa do KDD, e é um processo de descoberta de conhecimento em grandes quantidades de dados armazenados em *databases*, *data warehouses* ou outros tipos de repositórios.

2.2 Tarefas da mineração de dados

A descoberta de regras, padrões e conhecimentos das bases de dados estão ligados às tarefas da mineração de dados. A seguir um resumo das principais tarefas da mineração de dados (BERRY e LINOFF, 2004, p.8):

Classificação: consiste em examinar objetos ou registros, alocando-os em grupos ou classes previamente definidos segundo determinadas características.

Agrupamento: é uma tarefa de segmentação que consiste em dividir a população em grupos mais homogêneos chamados de subgrupos ou *clusters*. O que difere a tarefa de agrupamento da classificação é o fato de que no agrupamento os grupos não são previamente conhecidos.

Estimação/Predição: A estimação e predição são tarefas que consistem em definir um provável valor para uma ou mais variáveis.

Associação: é uma tarefa que consiste em determinar quais fatos ou objetos tendem a ocorrer juntos numa determinada transação. Por exemplo, numa compra em um supermercado, avaliar os itens que foram comprados juntos. Desta idéia é que surge o nome da técnica para tarefas de associação, denominado de *market basket analysis*.

Descrição: consiste em descrever de uma maneira simplificada e compacta a base de dados. A descrição, também conhecida como sumarização é uma tarefa que pode ser empregada numa etapa inicial das análises, proporcionando um melhor conhecimento da base.

2.3 Técnicas em mineração de dados

As técnicas utilizadas em mineração de dados têm sua origem basicamente em duas áreas: Inteligência Artificial (IA) e na Estatística.

A IA surgiu na década de 80, e as técnicas ligadas a esta área são: árvores de decisão, regras de indução e redes neurais artificiais (RNA).

Outras técnicas têm suas raízes na área da Estatística, assim como a Análise de Regressão, Análise Multivariada e dentro desta pode-se citar a Análise de Agrupamentos, Análise de Componentes Principais (ACP).

A escolha da técnica está ligada a tarefa da mineração de dados e ao tipo de variável. Resumidamente são apresentadas as principais técnicas e as respectivas tarefas mais apropriadas:

Árvores de decisão: técnica utilizada para tarefas de classificação.

Redes neurais artificiais: utilizada para tarefas de estimação, predição, classificação e agrupamentos. O algoritmo mais utilizado em redes neurais é o *backpropagation*.

Análise de cestas de compras (*Market Basket Analysis*): utilizada para criar regras de associação, revelar padrões existentes nos dados, fatos que tendem a ocorrer juntos.

Análise de regressão: é uma técnica que busca explicar uma ou mais variáveis de interesse (contínuas, ordinais ou binárias) em função de outras. Através de um modelo ajustado aos dados é uma técnica indicada para realizar predições.

Análise de agrupamentos: também conhecida por segmentação de dados, esta técnica é indicada para detectar a formação de possíveis grupos de uma base de dados. Dentre os algoritmos utilizados para detectar grupos tem-se: k-médias (*k-means), k-medoids*.

Análise de componentes principais: é uma técnica utilizada para reduzir a dimensionalidade de um conjunto de dados.

Mais detalhes sobre as técnicas e tarefas de mineração de dados pode ser encontrado em BERRY e LINOFF (2004); HAN e KAMBER (2001).

As técnicas de mineração se dividem ainda de acordo com o método de aprendizagem que, para FERNANDEZ (2003, p.7-8) podem ser:

- métodos de aprendizagem supervisionados
- métodos de aprendizagem não-supervsionados

Esta divisão está relacionada às tarefas da mineração.

Se a tarefa é realizar predição, classificação e estimação, os métodos de aprendizagem supervisionados são indicados, através da aplicação de técnicas, tais como, análise de regressão e redes neurais.

Porém, se a tarefa é realizar uma sumarização, agrupamento, associação, os métodos de aprendizagem não-supervisionados são indicados, através das técnicas de análise de componentes principais e análise de agrupamentos.

2.4 Pré-processamento dos dados

Segundo HAN e KAMBER (2001,p.108), o pré-processamento pode melhorar a qualidade dos dados, melhorando assim a acurácia e eficiência dos processos de mineração subseqüentes.

Sua aplicação é necessária, pois as bases de dados em geral são muito grandes (gigabytes ou mais) e contém registros que comprometem a qualidade dos dados, como por exemplo, registros inconsistentes, falta de informação (registros faltantes), registros duplicados, *outliers* (valores discrepantes), assimetria, transformação entre outros.

As técnicas de pré-processamento podem ser divididas em (HAN e KAMBER, 2001,p.105):

- Limpeza dos dados (*data cleaning*)
- Integração de dados (*data integration*)
- Transformação de dados (*data transformation*)
- Redução de dados (*data reduction*)

2.4.1 Limpeza dos dados

As rotinas para limpeza de dados consistem em uma investigação para detectar registros incompletos, duplicados e dados incorretos. Para corrigir os registros incompletos, chamados de valores faltantes ou *missing values,* algumas soluções são sugeridas, tais como:

- ignorar os registros;
- completar manualmente os valores faltantes;
- substituir por uma constante global;
- uso da média para preencher os valores faltantes;
- uso do valor mais provável, que pode ser predito com auxílio de uma regressão, árvores de decisão, entre outras.

Outro problema tratado nesta etapa do pré-processamento é a presença de *outiliers*. *Outilers* são dados que possuem um valor atípico para uma determinada variável, ou com características bastante distintas dos demais. Estes valores podem ser detectados através da análise de agrupamento, onde os valores similares formam grupos, destacando-se os *outliers* (HAN e KAMBER, 2001,p.111).

As soluções citadas para tratar valores *missing,* também podem ser adotadas para *outilers*, porém a exclusão do valor só deve ser realizada quando o dado representar um erro de observação, de medida ou algum problema similar.

Tanto *missing values* quanto *outliers* precisam receber um tratamento para que não comprometam as análises seguintes.

2.4.2. Integração dos dados

Freqüentemente, para se fazer a mineração de dados, é necessário integrar dados armazenados em diferentes fontes ou bases. A integração pode ser feita combinando variáveis que estão em diferentes bases. Aproximadamente, 70% do tempo gasto na mineração é devido à preparação dos dados obtidos de diferentes bases (FERNANDEZ, 2003, p.6,15).

2.4.3 Transformação dos dados

Segundo HAN e KAMBER (2001, p.114) e HAYKIN (1999, p.205,208), a transformação das variáveis originais pode melhorar a eficiência dos algoritmos de classificação envolvendo redes neurais. Dentre os tipos de transformação, os autores citam a normalização min-max, definida por: Seja uma determinada variável A, com valores $A_1, A_2, A_3, ..., A_n$ .Sendo o valor mínimo de A representado por $\min_A$ e o valor máximo de A por $máx_A$ e deseja-se transformar os valores $A_1, A_2, A_3, ..., A_n$ para valores em um intervalo [a,b], então os novos valores $A_1', A_2', A_3', ..., A_n'$ são dados pela equação:

$$A_i' = \frac{A_i - \min_A}{\max_A - \min_A}.(b-a) + a \qquad \text{i = (1,2,3, ..., n)} \qquad (2.1)$$

A transformação de variáveis também pode auxiliar as técnicas estatísticas que se baseiam na suposição da normalidade dos dados. Em muitas situações práticas ocorre que a distribuição dos dados é assimétrica. Uma transformação da variável pode aproximar os dados de uma distribuição normal, tornando-a mais simétrica. As transformações mais exploradas para este caso são (BUSSAB e MORETTIN, 2003, p.53):

$$x^p = \begin{cases} x^p, & se \quad p > 0 \\ \ln(x), & se \quad p = 0 \\ -x^p & se \quad p < 0 \end{cases}$$

sendo *p* escolhido na seqüência ...,-3,-2,-1,-1/2 , -1/3, -1/4, 0, ¼, 1/3, ½, 1, 2, 3, ... Para cada valor de *p* obtém-se gráficos apropriados (histogramas, *Box Plots*) para os dados originais e transformados, de modo a escolher o valor mais adequado de *p*.

2.4.4 Redução da dimensionalidade

Uma das razões para se realizar uma redução da dimensionalidade está no tempo computacional que os algoritmos podem levar para realizar uma tarefa de mineração de

dados. Apesar de que o tempo vai depender de fatores, tais como: o tamanho da base (número de variáveis e observações), a tarefa a ser realizada e o algoritmo (se é iterativo ou não), numa base reduzida o seu desempenho tende a ser mais rápido.

Entre as estratégias aplicadas na redução, HAN e KAMBER (2001,p.121) citam a técnica de componentes principais. Esta técnica estatística é bem conhecida e muitas vezes aplicada como etapa de pré-processamento em grandes bases de dados (JOHNSON e WICHERN, 2002, p.426).

2.5 Importância da redução da dimensionalidade

Considerando uma situação em que há muitas observações (casos, registros) e variáveis, como é possível entender o relacionamento destas variáveis? Quais os casos similares que podem ser agrupados? Existem valores discrepantes (*outliers)*? Quais?
Estas perguntas para serem respondidas, quando se tem uma grande base de dados não é uma tarefa simples. Trabalhando-se com um grande número *m* de variáveis, o ideal seria obter uma representação num espaço de menor dimensionalidade, onde ficaria mais simples observar as relações. Esta redução pode ser importante a fim de obter um melhor aproveitamento dos dados, facilitando a extração de conhecimento, através de uma representação gráfica ou gerando um novo conjunto de coordenadas, menor que a original, para utilizá-lo em outras tarefas de mineração de dados com alto tempo computacional.

Por outro lado, ao fazer a redução do espaço real de representação corre-se o risco de perder informação. Esta perda de informação significa que o conjunto reduzido não revela de forma confiável características presentes nos dados originais.

Uma técnica estatística muito conhecida e aplicada, que permite a redução de variáveis quantitativas, estudada por Pearson em 1901 e consolidada por Hotteling em 1933, é chamada análise de componentes principais (ACP) (LEBART *et al*., 1995, p.32).
Esta técnica multivariada consiste, basicamente, em analisar um conjunto de variáveis numéricas com alta dimensionalidade de representação, reduzindo o número de variáveis, mantendo a máxima variabilidade dos dados originais, minimizando assim a perda de informação ao se fazer redução.

2.6 Componentes principais e suas aplicações

A análise de componentes principais freqüentemente tem sido utilizada como uma etapa intermediária de grandes análises, podendo servir como pré-processamento para outras técnicas, como por exemplo, regressão múltipla e análise de agrupamentos (JOHNSON e WICHERN , 2002, p.426).

Segundo MANLY (2005, p.130), alguns algoritmos de análise de agrupamentos iniciam através da análise de componentes principais, reduzindo as variáveis originais num número menor de componentes. Porém, os resultados de uma análise de agrupamentos podem ser bem diferentes se for utilizado inicialmente a ACP, uma vez que ao reduzir a dimensão dos dados pode-se perder informação. Outro aspecto a ser levado em consideração é em relação à interpretação das componentes, uma vez que seu significado não é claro como as variáveis originais. Para usar a ACP, o ideal é que a variabilidade dos dados seja explicada por poucas componentes, tornando assim a redução um caminho útil para uma análise de agrupamentos, se esta for a tarefa de mineração de dados a ser realizada.

A ACP como etapa de pré-processamento para outras tarefas também é indicada como método para melhorar o desempenho do algoritmo de retropropagação em Redes Neurais (RN). O algoritmo de retropropagação ou *backpropagation* é um algoritmo de aprendizagem muito utilizado em RN. Segundo HAYKIN (1999, p.208), este algoritmo tem sua performance melhorada se as variáveis não forem correlacionadas. Esta tarefa de gerar variáveis não correlacionadas pode ser executada pela ACP. Neste caso, a preocupação maior é de gerar as variáveis não correlacionadas, dadas pelas componentes, podendo ser utilizadas todas as componentes, caso necessário evitar perda de informação.

Outra aplicação da ACP está relacionada à análise de regressão, em seu uso como uma alternativa para obter pelo método de mínimos quadrados os coeficientes de regressão na presença de multicolinearidade, sendo uma técnica eficiente para detectar a multicolinearidade (CHATTERJEE *et al*., 2000, p.269).

A ACP também tem muitas aplicações na Engenharia de Produção, como exemplo determinar a confiabilidade e o tempo médio de falha de peças em um equipamento, auxiliando as empresas (SCREMIN, 2003, p. 39). Em seu trabalho, SCREMIN (2003,

p.41) relata um estudo feito sobre propriedades rurais do Estado de Santa Catarina. O estudo envolveu 27 variáveis, reduzidas a 6 componentes principais, utilizando em seguida Redes Neurais para identificar grupos. O resultado, segundo o autor, foi bem positivo, onde foi possível identificar quatro grupos homogêneos.

2.7 Componentes principais na mineração de dados

Além das aplicações de componentes principais como pré-processamento para outras técnicas estatísticas, a ACP também é muito utilizada na mineração de dados. Foram consultados alguns artigos recentes sobre o assunto. Dentre os artigos, pode-se citar:

. *Association Rule Discovery in Data Mining by Implementing Principal Component Analysis* (GERARDO *et. al*, 2004)*;*

. *Distributed Clustering Using Collective Principal Component Analysis* (KARGUPTA *et al*., 2001)*;*

. *Outlier Mining Based on Principal Component Estimation* (YANG e YANG, 2004).

Os resumos dos artigos consultados estão disponíveis no endereço: <www.springerlink.com>.

Com base nos resumos desses artigos, a ACP é utilizada na mineração de dados para:

- descoberta de regras de associação, através da implementação de várias técnicas.
- obtenção de subgrupos ou *clusters*.
- detectar *outilers.*

Na mineração de dados pode ser útil reduzir a dimensionalidade da base, principalmente quando se deseja ganhar tempo de processamento ou mesmo para simplificar a base de dados.

No capítulo 5 será apresentado o pré-processamento utilizando ACP com escalonamento ótimo como etapa preliminar da mineração de dados.

No capítulo 6, exemplificando o uso da ACP com escalonamento ótimo como etapa de pré-processamento na mineração de dados, será utilizada uma base de dados do IBGE, com 118.76 registros e variáveis mensuradas em diferentes níveis. Na seqüência será feito um estudo da base, com o objetivo de descrever grupos similares de pessoas, através da Análise de Agrupamentos.

3. COMPONENTES PRINCIPAIS

Este capítulo tem como objetivo fazer a revisão de literatura em componentes principais (ACP), incluindo exemplos. O desenvolvimento aqui apresentado baseia-se em JOHNSON e WICHERN (2002), BANET e MORINEAU (1999), MANLY (2005) e REIS (1997).

A revisão não está centrada em demonstrações matemáticas, mas sim na compreensão da técnica por meio de exemplos. A ACP conduz à obtenção de um novo conjunto de coordenadas, menor que o original, a fim de ser utilizado para descrever os dados, ou mesmo para ser utilizado em outras técnicas de análise ou de mineração de dados. As representações gráficas aqui apresentadas foram obtidas através dos *softwares* Statistica 6.0 e LHSTAT (LOESCH, 2005).

3.1 Análise de componentes principais (ACP)

Analisar conjuntos de dados em espaços com alta dimensionalidade para obter conclusões claras e confiáveis não é uma tarefa fácil. Uma técnica estatística, chamada componentes principais, criada por Karl Pearson em 1901, e posteriormente consolidada por Harold Hottelling em 1933, continua sendo muito utilizada em diversas áreas do conhecimento.

Segundo JOHNSON e WICHERN (2002, p.426), a técnica de componentes principais tem como objetivo geral a redução da dimensionalidade e interpretação de um conjunto de dados. Obter esta redução num conjunto de *m* variáveis $(X_1, X_2, X_3, ..., X_m)$, consiste em encontrar combinações lineares das *m* variáveis, que irão gerar um outro conjunto de variáveis $(Y_1, Y_2, Y_3, ..., Y_m)$ com novas coordenadas e não correlacionadas entre si.

Geometricamente, as componentes principais representam um novo sistema de coordenadas, obtidas por uma rotação do sistema original, que fornece as direções de

máxima variabilidade, e proporciona uma descrição mais simples e eficiente da estrutura de covariância dos dados.

Para BANET e MORINEAU (1999, p.15) e JOHNSON e WICHERN (2002, p.426), a análise de componentes principais pode ser uma etapa intermediária de cálculo para uma análise posterior, como: regressão múltipla, análise de agrupamentos, classificação, discriminação, redes neurais, entre outras. Assim, a aplicação de componentes principais sob um conjunto de dados, poderá ser extremamente útil, a fim de gerar soluções para uma classe de problemas em mineração de dados que exige a redução de dimensionalidade, como uma etapa de pré-processamento.

3.2 Representação da matriz de dados e a nuvem de pontos

Seja um conjunto de dados com *m* variáveis numéricas, $X_1, X_{2,} X_3, ...X_m$, com *n* observações ou casos, representados na Tabela 3.1.

Tabela 3.1 Conjunto de dados : *n* observações e *m* variáveis

Observação	X_1	X_2	...	X_m
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nm}

Este conjunto de dados pode ser escrito em forma de uma matriz, que terá *n* linhas por *m* colunas e pode ser genericamente representado por :

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ ... & ..... & .... & ..... \\ x_{n1} & x_{n2} & .... & x_{nm} \end{bmatrix},$$

em que cada elemento $\mathbf{x}_{ij}$ desta matriz é um número conhecido que representará a observação i (i = 1, 2, 3, ... ,n) em termos da variável j (j = 1, 2, 3, ... ,m).

Exemplo 3.1: Considere um conjunto de dados em que foram pesquisadas 8 pessoas em relação a três variáveis: X_1 :peso (kg), X_2 :altura (cm) e X_3 :idade (anos). Os dados são apresentados na Tabela 3.2

Tabela 3.2 Conjunto de dados com 8 observações e 3 variáveis

	X_1	X_2	X_3
Observação	Peso	Altura	Idade
1	55	164	25
2	90	185	18
3	79	179	47
4	60	172	45
5	83	177	49
6	82	176	50
7	95	189	65
8	54	160	23

Na forma matricial, tem-se: $\mathbf{X} = \begin{bmatrix} 55 & 164 & 25 \\ 90 & 185 & 18 \\ 79 & 179 & 47 \\ 60 & 172 & 45 \\ 83 & 177 & 49 \\ 82 & 176 & 50 \\ 95 & 189 & 65 \\ 54 & 160 & 23 \end{bmatrix}$,

Portanto, $\mathbf{X}_{n \times m}$, com n = 8 observações e m = 3 variáveis.

Identificar visualmente indivíduos semelhantes em uma matriz desse tipo, não é tão simples quando se está diante de um número elevado de observações e variáveis. Considerando cada linha da matriz de dados **X** (neste exemplo, informações de uma pessoa), como um ponto de coordenadas nas variáveis consideradas (peso, altura e idade), pode-se utilizar um espaço de *m* dimensões para representar esta matriz, formando uma nuvem de *n* pontos, o que BANET e MORINEAU (1999, p.20), chamam nuvem de pontos-linha.

A representação gráfica dos pontos-linha é dada na Figura 3.1.

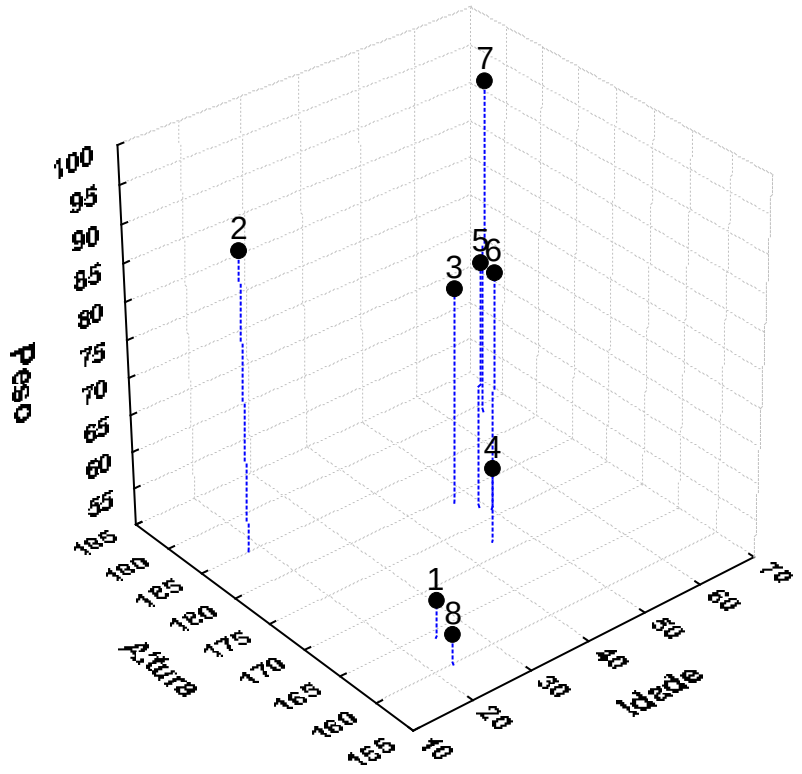

Figura 3.1 Representação gráfica da matriz de dados da Tabela 3.2 formando a nuvem de pontos-linha.

Neste exemplo, a nuvem de pontos-linha é formada por 8 observações, em um espaço tridimensional. A partir da nuvem de pontos pode-se fazer algumas considerações. Os pontos que aparecem próximos um do outro, indicam indivíduos ou pessoas, com características semelhantes em cada uma das três variáveis. Assim, os indivíduos 3, 5 e 6 têm medidas semelhantes nas três variáveis, o mesmo ocorre com os indivíduos 1 e 8. Os indivíduos mais distantes, indicam que suas medidas, diferem entre si, em pelo menos uma variável, como é o caso do indivíduo 7.

De forma análoga, pode-se considerar cada coluna da matriz de dados **X**, representada em um espaço de dimensão n (neste exemplo n = 8). Esta representação é chamada de nuvem de pontos-coluna ou pontos-variáveis (BANET e MORINEAU, 1999, p.21). Os pontos-variáveis próximos indicarão variáveis correlacionadas em termos do conjunto de dados considerado. Porém, a representação gráfica dos pontos-

variáveis para este exemplo nem é possível, uma vez que são vetores de dimensão n = 8 ($\mathbb{R}^8$).

As representações dessas nuvens em seus espaços respectivos estabelecem uma dualidade entre as colunas e as linhas da matriz de observações. Esta dualidade é de relevante importância na análise de componentes principais, pois é através dela que é possível entender o relacionamento entre as variáveis, entre as observações, e entre as observações e variáveis.

No exemplo 3.1 foi possível representar visualmente as 8 observações, utilizando os eixos para as 3 variáveis. Porém, se fosse incluída mais uma variável, esta representação visual não seria mais possível. A essência da análise de componentes principais é tornar compreensível estas nuvens de pontos que se encontram em espaços de dimensão elevada, através da busca de um subespaço, chamado subespaço vetorial, sobre o qual projetamos a nuvem de pontos original. O novo conjunto de coordenadas deverá ser o mais parecido possível com a configuração da nuvem original. Quando o subespaço tem apenas duas dimensões, recebe o nome de plano fatorial. A Figura 3.2 ilustra esta idéia.

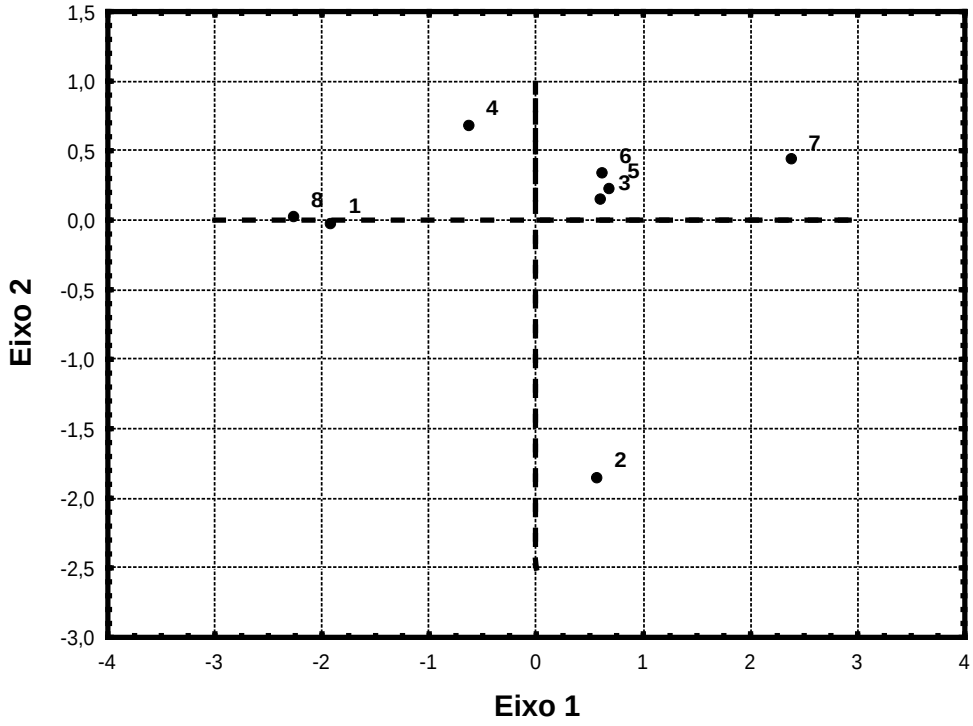

Figura 3.2 Plano fatorial de duas dimensões projetando os pontos-linha

Através de uma análise dos dados da Tabela 3.2 e da Figura 3.1, pode-se observar que o plano fatorial de projeção da Figura 3.2 reproduz, aproximadamente, em 2 dimensões, a configuração originalmente apresentada em m = 3 dimensões.

Esta redução no conjunto de coordenadas permitirá um entendimento das relações entre as observações, entre as variáveis e entre observações e variáveis. Além da possibilidade de representação gráfica, as novas coordenadas das direções 1 e 2 também podem servir como um novo conjunto de dados para outras técnicas serem aplicadas.

3.3 Subespaço vetorial de k dimensões

Como os pontos-linha estão no espaço $\mathbb{R}^m$, m dimensões serão necessárias para a sua representação. A partir dos dados da Tabela 3.2, a representação dos pontos-linha seria num espaço de m = 3 dimensões, conforme mostrado na Figura 3.1.

Ao reduzir a representação para um subespaço de k = 2 dimensões, conforme a Figura 3.2, encontra-se um subespaço $\mathbb{R}^k$, com k < m, de forma que este novo subespaço reproduz o mais próximo possível a representação original.

Para BANET e MORINEAU (1999, p.25), o plano de projeção que contém a máxima dispersão entre os pontos é também o plano que se aproxima o máximo possível da nuvem original. Prova-se, matematicamente, que a maximização da variabilidade das projeções da nuvem equivale à minimização da soma dos quadrados das distâncias entre os pontos da nuvem e o subespaço.

A Figura 3.3 apresenta um subespaço de k = 2 dimensões para representação dos pontos-linha. Os vetores de comprimento unitário $\mathbf{v}_1$ e $\mathbf{v}_2$, nas direções dos eixos 1 e 2, respectivamente , constituem uma base deste subespaço, que vem a ser um plano.

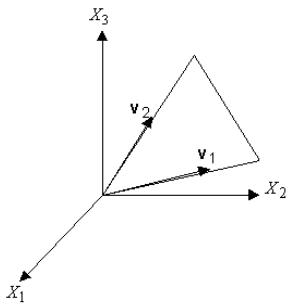

Figura 3.3 Subespaço com 2 dimensões determinado pelos vetores $\mathbf{v}_1$ e $\mathbf{v}_2$

A direção $\mathbf{v}_1 = (v_{11},...,v_{1m})^t$, obtida de tal maneira que a variabilidade dos pontos projetados sobre ela seja máxima, gera a primeira componente principal: $Y_1 = v_{11}\mathbf{X}_1 + ... + v_{1m}\mathbf{X}_m$; a direção $\mathbf{v}_2 = (v_{21},...,v_{2m})^t$, que possui a segunda maior variabilidade, com restrição a ser ortogonal a $\mathbf{v}_1$, gera a segunda componente principal: $Y_2 = v_{21}\mathbf{X}_1 + ... + v_{2m}\mathbf{X}_m$.

Os vetores $\mathbf{v}_1$ e $\mathbf{v}_2$ pertencentes ao $\mathbb{R}^m$ devem satisfazer :

1) São perpendiculares entre si : $\mathbf{v}_1^t\ \mathbf{v}_2 = 0$;
2) Possuem comprimento unitário : $\mathbf{v}_1^t\ \mathbf{v}_1 = \mathbf{v}_2^t\ \mathbf{v}_2 = 1$;
3) Constituem uma base para o subespaço vetorial.

3.4 Obtenção das componentes principais

Para obter as componentes principais, uma maneira é padronizar os dados e calcular a matriz de correlações, cujos passos são apresentados na seqüência:

1º - Padronização dos dados da matriz **X**

A padronização dos dados da matriz **X** irá gerar uma matriz **Z**, cujos elementos z_{ij} são calculados pela expressão :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (3.1)$$

em que $\bar{x}_j$ é a média da coluna (variável) j e s_j o desviopadrão amostral da coluna (variável) j, na matriz **X**.

No exemplo 3.1, após a aplicação da expressão (3.1), obtém-se:

$$\mathbf{Z} = \begin{bmatrix} -1{,}225 & -1{,}149 & -0{,}931 \\ 0{,}946 & 0{,}995 & -1{,}359 \\ 0{,}263 & 0{,}383 & 0{,}412 \\ -0{,}915 & -0{,}332 & 0{,}290 \\ 0{,}512 & 0{,}179 & 0{,}534 \\ 0{,}500 & 0{,}077 & 0{,}595 \\ 1{,}256 & 1{,}404 & 1{,}511 \\ -1{,}287 & -1{,}557 & -1{,}053 \end{bmatrix}$$

2º - Calculo da matriz de correlações amostrais, **R**

A matriz de correlações amostrais, **R**, sob os dados padronizados, é dada por:

$$\mathbf{R} = \frac{1}{n-1}\left(\mathbf{Z}^t\mathbf{Z}\right) \tag{3.2}$$

No exemplo 3.1, $\mathbf{R} = \begin{bmatrix} 1{,}000 & 0{,}949 & 0{,}499 \\ 0{,}949 & 1{,}000 & 0{,}526 \\ 0{,}499 & 0{,}526 & 1{,}000 \end{bmatrix}$

3º - Cálculo dos autovalores

Os autovalores, $\lambda_1, \lambda_2, \ldots, \lambda_m$, são calculados sob a matriz de correlação, **R**, de dimensão m x m, segundo a equação característica:

$$\left| \mathbf{R} - \lambda\mathbf{I} \right| = 0 \tag{3.3}$$

onde **I** é a matriz identidade de ordem m..

Esta equação tem solução matemática. Existem exatamente m autovalores, não negativos: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq 0$; e correspondentes aos autovalores, também m autovetores, $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m$.

No exemplo 3.1, após a aplicação da equação (3.3), obtém-se:

$$\left| \begin{bmatrix} 1{,}000 & 0{,}949 & 0{,}499 \\ 0{,}949 & 1{,}000 & 0{,}526 \\ 0{,}499 & 0{,}526 & 1{,}000 \end{bmatrix} - \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right| = 0$$

A resolução da equação acima resulta nos seguintes autovalores:

$\lambda_1 = 2{,}3412$; $\lambda_2 = 0{,}6088$ e $\lambda_3 = 0{,}0500$

Uma propriedade importante é que a soma dos autovalores é igual ao traço da matriz de correlações, portanto igual a m. No exemplo apresentado, tem-se que 2,3412 + 0,6088 + 0,0500 = 3.

4º - Cálculo dos autovetores

Os autovetores, $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m$, podem ser calculados segundo a equação:

$$\mathbf{Ru} = \lambda \mathbf{u} \qquad (3.4)$$

No exemplo 3.1, após a aplicação da equação (3.4), obtém-se

$$\begin{bmatrix} 1{,}000 & 0{,}949 & 0{,}499 \\ 0{,}949 & 1{,}000 & 0{,}526 \\ 0{,}499 & 0{,}526 & 1{,}000 \end{bmatrix} . \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix} = \lambda . \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}$$

A solução da equação acima resulta nos m autovetores $\mathbf{U}_j$, (j = 1, 2, ..., m) não normalizados.

5º - Normalização dos autovetores

A normalização dos autovetores $\mathbf{U}_j$ é dado pela expressão:

$$\mathbf{v}_j = \frac{1}{\sqrt{\mathbf{u}_j^t \mathbf{u}_j}} \mathbf{u}_j \qquad (3.5)$$

Aplicando (3.5), os autovetores normalizados do exemplo 3.1 são:

$$\mathbf{v}_1 = \begin{bmatrix} 0{,}61928 \\ 0{,}62484 \\ 0{,}47547 \end{bmatrix},\ \mathbf{v}_2 = \begin{bmatrix} -0{,}3572 \\ -0{,}3150 \\ 0{,}8793 \end{bmatrix} \text{ e } \mathbf{v}_3 = \begin{bmatrix} 0{,}6992 \\ -0{,}7144 \\ 0{,}0281 \end{bmatrix}$$

Os m autovetores normalizados podem ser escritos como uma matriz $\mathbf{V}_{m\,x\,m}$, sendo que esses autovetores representam os coeficientes das componentes principais. Assim, tem-se a matriz:

$$\mathbf{V} = \begin{bmatrix} 0{,}61928 & -0{,}3572 & 0{,}6992 \\ 0{,}62484 & -0{,}3150 & -0{,}7144 \\ 0{,}47547 & 0{,}8793 & 0{,}0281 \end{bmatrix}$$

com as componentes principais dadas segundo as equações :

$$Y_1 = 0{,}61928.Z_1 + 0{,}62484.Z_2 + 0{,}47547.Z_3$$

$$Y_2 = -0{,}3572.Z_1 - 0{,}3150.Z_2 + 0{,}8793.Z_3$$

$$Y_3 = 0{,}6992.Z_1 - \ 0{,}7144.Z_2 + 0{,}0281.Z_3$$

onde $Z_j = \frac{X_j - \overline{X}_j}{S_j}$, conforme definido na expressão (3.1), refere-se a variável padronizada.

O número de componentes principais será igual ao número de variáveis utilizadas na análise e todas as componentes principais explicam cem por cento da variabilidade original.

3.5 Dados em subgrupos

Muitas vezes os dados estão divididos em subgrupos ou estratos conhecidos. Isto acontece, por exemplo, numa tarefa de classificação na mineração de dados. Neste caso, sendo g o número de subgrupos, a ACP deve ser realizada da seguinte maneira (GNANADESIKAN, 1977, p.12):

$$\mathbf{S} = \frac{1}{n-g}\sum_{g=1}^{g}\left(n_g - 1\right)\mathbf{S}_g \qquad (3.6)$$

em que:
n: número total de casos;
g: número de subgrupos;
n_g : número de casos no subgrupo g;
$\mathbf{S}_g$: Matriz de covariância calculada para o subgrupo g.

A padronização da matriz de covariâncias S equivale a matriz de correlações R, calculadas segundo as expressões (3.1) e (3.2).

3.6 Variâncias das CP´s

A soma dos autovalores da matriz de correlação, **R**, é igual ao número de variáveis na análise, e este valor é chamado de variância total ou inércia. Então, a variância total é igual a $\lambda_1 + \lambda_2 + ... + \lambda_m$. Esta inércia é a mesma para a nuvem de pontos-linha e pontos-coluna (BANET e MORINEAU, 1999, p.34), e $\lambda_j = Var\left(Y_j\right)$ (j=1, 2, 3, ...,m).

Assim, a variância total do sistema para o exemplo 3.1 é igual a 3 e a contribuição de cada componente pode ser calculada por:

$$\frac{\lambda_j}{m},$$

(3.7)

sendo m o número de variáveis do modelo.

Do exemplo 3.1, tem-se que a primeira componente, Y_1, explica $\frac{2,3412}{3} = 0,7804 = 78,04\%$ da variância total em apenas uma dimensão.

Geralmente, ordenam-se os m autovalores de forma que $\lambda_1 \geq \lambda_2, \geq ... \geq \lambda_m$ e escolhem-se os k maiores autovalores λ_1 ,..., λ_k . Seus correspondentes autovetores $\mathbf{v}_1$, ..., $\mathbf{v}_k$ são os que determinam as direções principais. Ainda tem-se que a inércia total pode ser decomposta em dois tipos: explicada e não explicada. A inércia explicada está relacionada ao poder de explicação ao se fazer a redução de m dimensões para k dimensões (k < m) e a inércia não explicada está relacionada à perda de explicação devido a redução no conjunto de coordenadas.

$$\underbrace{tr(\mathbf{R})}_{\text{Inércia Total}} = \underbrace{\lambda_1 + \lambda_2 + ... + \lambda_k}_{\text{Inércia Explicada}} + \underbrace{\lambda_{k+1} + ... + \lambda_m}_{\text{Inércia não Explicada}}. \quad (3.8)$$

No exemplo 3.1, em que m=3, se reter λ_1 e λ_2, tem-se que:

Inércia explicada ou variância explicada: $\lambda_1 + \lambda_2 = 2,3412 + 0,6088 = 2,95$;

Inércia não-explicada: $\lambda_3 = 0,05$.

Percentual da inércia explicada : $\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{2,95}{3} = 0,9833 = 98,33\%$.

Isto significa que 98,33% da inércia total é explicada pelos pontos projetados no plano fatorial de duas dimensões, ou seja, retendo duas componentes é possível explicar 98,33% da variabilidade original dos dados.

3.7 Número de componentes principais

Existem vários critérios práticos para determinar o número ideal de componentes principais a considerar. Na seqüência serão apresentados os mais conhecidos (detalhes em REIS,1997, p.272).

3.7.1 Critério do *Scree plot*

Proposto por Cattell em 1966, consiste em representar a porcentagem de variância explicada por cada componente, conforme o gráfico da Figura 3.4 o qual foi obtido a partir de dados hipotéticos. Quando esta porcentagem se reduz e a curva passa a ser quase paralela ao eixos das abscissas, os componentes seguintes deverão ser excluídos.

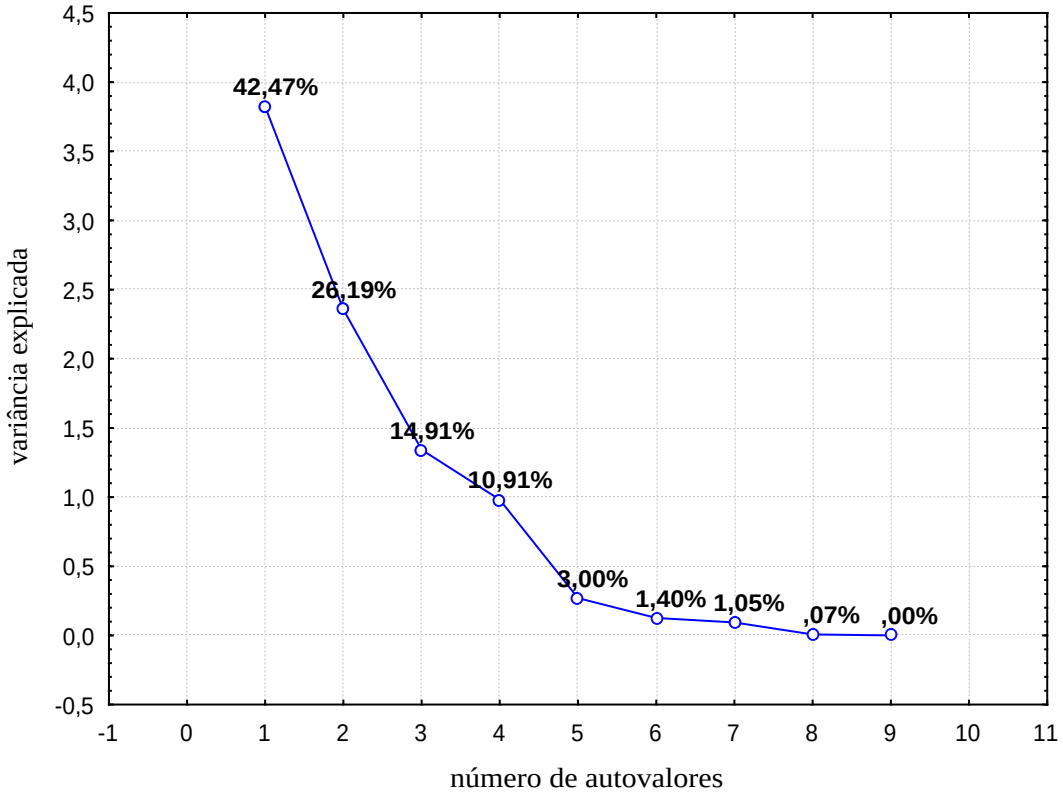

Figura 3.4 : *Scree plot*

No exemplo da Figura 3.4, a curva passa a ser quase paralela ao eixo das abscissas a partir do quinto ou sexto autovalor. Então, segundo este critério dever-se-ia optar por 5 ou 6 componentes.

3.7.2 Critério de Kaiser

Segundo este critério, deve-se excluir as componentes cujos autovalores são inferiores à média aritmética de todos os autovalores, isto é, menores do que 1 se a análise for feita a partir da matriz de correlações. A Tabela 3.3 apresenta os autovalores para o exemplo 3.1.

Tabela 3.3 Autovalores do exemplo 3.1

λ_i	Autovalor	Variância Explicada	Variância Explicada (acumulada)
1	2,34	78,04	78,04
2	0,61	20,29	98,33
3	0,05	1,67	100,00
Total	3	100	

Segundo este critério deve-se reter apenas a primeira componente principal, já que o segundo autovalor apresenta valor igual a 0,61, menor do que 1.

3.7.3 Critério baseado na porcentagem acumulada da variância explicada

Este critério sugere considerar o número de componentes suficientes para explicar mais de 70% da variância total. Observando a Tabela 3.3, é possível concluir que apenas a primeira componente principal explica mais de 70% da variação.

3.7.4 Critério baseado na lógica difusa

Este critério para a seleção de componentes principais é apresentado por SCREMIN (2003). O autor apresenta um critério baseado na lógica difusa, em que a seleção do número de componentes principais contemple as variâncias explicadas pelos fatores, as porcentagens acumuladas de variância explicada, as cargas fatoriais e o conhecimento do pesquisador e/ou especialista sobre o problema, descritos em forma de atributos lingüísticos.

3.8 Projeção de um ponto e as novas coordenadas

Seja um subespaço com duas dimensões, representado na Figura 3.5, em que $\mathbf{v}_1$ e $\mathbf{v}_2$ determinam as duas direções principais, $\hat{x}_i$ é a projeção perpendicular de x_i sobre o plano de duas dimensões.

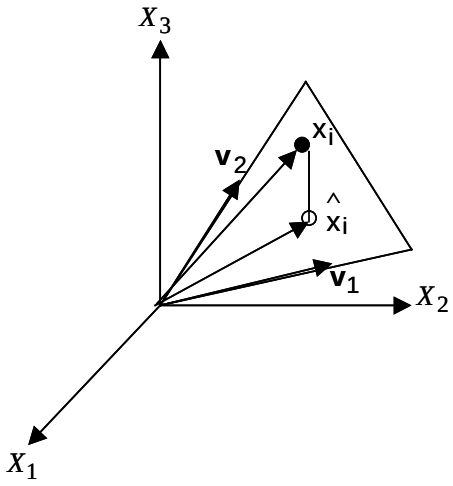

Figura 3.5 - Projeção de um ponto x_i sobre o plano fatorial.

Sejam y_{1i} e y_{2i} as coordenadas de $\hat{x}_i$ em relação à base $\{\mathbf{v}_1, \mathbf{v}_2\}$. Se uma determinada linha i possui valores x_{i1}, x_{i2}, x_{i3}, ..., x_{im}, cujos valores padronizados são z_{i1}, z_{i2}, z_{i3}, ..., z_{im} calculam-se suas coordenadas y_{1i} (na direção $\mathbf{v}_1$) e y_{2i} (na direção $\mathbf{v}_2$), segundo as operações vetoriais:

$$y_{1i} = z_{i1}v_{11} + z_{i2}v_{21} + z_{i3}v_{31} + ... + z_{im}v_{n1} \quad (3.9)$$

$$y_{2i} = z_{i1}v_{12} + z_{i2}v_{22} + z_{i3}v_{32} + ... + z_{im}v_{n2} \quad (3.10)$$

onde cada elemento z_{ij} é calculado segundo a expressão (3.1).

Dada a matriz **V** formada pelos autovetores normalizados, considerando apenas as duas primeiras colunas de **V**, que representam as duas direções principais; e a matriz **Z** que contém os dados padronizados do exemplo 3.1.

Para o primeiro caso, i = 1, tem-se:

$$\mathbf{V} = \begin{bmatrix} 0{,}61928 & -0{,}3572 \\ 0{,}62484 & -0{,}3150 \\ 0{,}47547 & 0{,}8793 \end{bmatrix} \qquad \text{e } \mathbf{Z}_1 = \begin{bmatrix} -1{,}225 & -1{,}149 & -0{,}931 \end{bmatrix}$$

Aplicando as expressões (3.9) e (3.10) e considerando apenas duas direções principais, obtém-se as novas coordenadas (y_{11}, y_{21}) para o primeiro ponto-linha. Utilizando os dados padronizados, têm-se:

$$y_{11} = (\text{-}1{,}225)(0{,}61928) + (\text{-}1{,}149)(0{,}62484) + (\text{-}\,0{,}931)(0{,}47547) = \text{ -}1{,}919$$
$$y_{21} = (\text{-}1.225)(\text{-}0{,}3572) + (\text{-}1{,}149)(-0{,}3150) + (\text{-}\,0{,}931)(0{,}8793) = \text{ -}\,0{,}019$$

Logo, as novas coordenadas deste primeiro ponto-linha serão (-1,919; -0,019).

Segundo as expressões (3.9) e (3.10), as novas coordenadas podem ser calculadas para as demais observações, i = 2, 3, ...,8. O resultado da aplicação das expressões (3.9) e (3.10) determina um novo conjunto de coordenadas, representado pela matriz **Y**, que matricialmente pode ser obtido por:

$$\mathbf{Y} = \mathbf{ZV} \tag{3.11}$$

No exemplo 3.1, aplicando a expressão (3.11), tem-se que:

$$\mathbf{Y} = \begin{bmatrix} -1{,}225 & -1{,}149 & -0{,}931 \\ 0{,}946 & 0{,}995 & -1{,}359 \\ 0{,}263 & 0{,}383 & 0{,}412 \\ -0{,}915 & -0{,}332 & 0{,}290 \\ 0{,}512 & 0{,}179 & 0{,}534 \\ 0{,}500 & 0{,}077 & 0{,}595 \\ 1{,}256 & 1{,}404 & 1{,}511 \\ -1{,}287 & -1{,}557 & -1{,}053 \end{bmatrix} \begin{bmatrix} 0{,}61928 & -0{,}3572 & 0{,}6992 \\ 0{,}62484 & -0{,}3150 & -0{,}7144 \\ 0{,}47547 & 0{,}8793 & 0{,}0281 \end{bmatrix} = \begin{bmatrix} -1{,}919 & -0{,}019 & -0{,}062 \\ 0{,}562 & -1{,}846 & -0{,}088 \\ 0{,}598 & 0{,}148 & -0{,}078 \\ -0{,}636 & 0{,}686 & -0{,}394 \\ 0{,}683 & 0{,}231 & 0{,}245 \\ 0{,}609 & 0{,}339 & 0{,}276 \\ 2{,}374 & 0{,}438 & -0{,}082 \\ -2{,}271 & 0{,}024 & 0{,}183 \end{bmatrix}$$

Considerando trabalhar apenas com duas dimensões, k = 2, escolhe-se apenas as duas primeiras colunas da matriz **Y**, obtendo-se, assim, uma redução na dimensionalidade, de m=3 para k = 2, preservando 98,33% da variância total.

A matriz $\mathbf{Y}_{8x2}$ apresenta uma redução de dimensionalidade, podendo ser utilizada em análises posteriores (mineração de dados). O resultado de $\mathbf{Y}_{8x2}$ é dado por:

$$\mathbf{Y} = \begin{bmatrix} -1{,}919 & -0{,}019 \\ 0{,}562 & -1{,}846 \\ 0{,}598 & 0{,}148 \\ -0{,}636 & 0{,}686 \\ 0{,}683 & 0{,}231 \\ 0{,}609 & 0{,}339 \\ 2{,}374 & 0{,}438 \\ -2{,}271 & 0{,}024 \end{bmatrix},$$

e a representação das oito observações nas novas coordenadas é dada pela Figura 3.6.

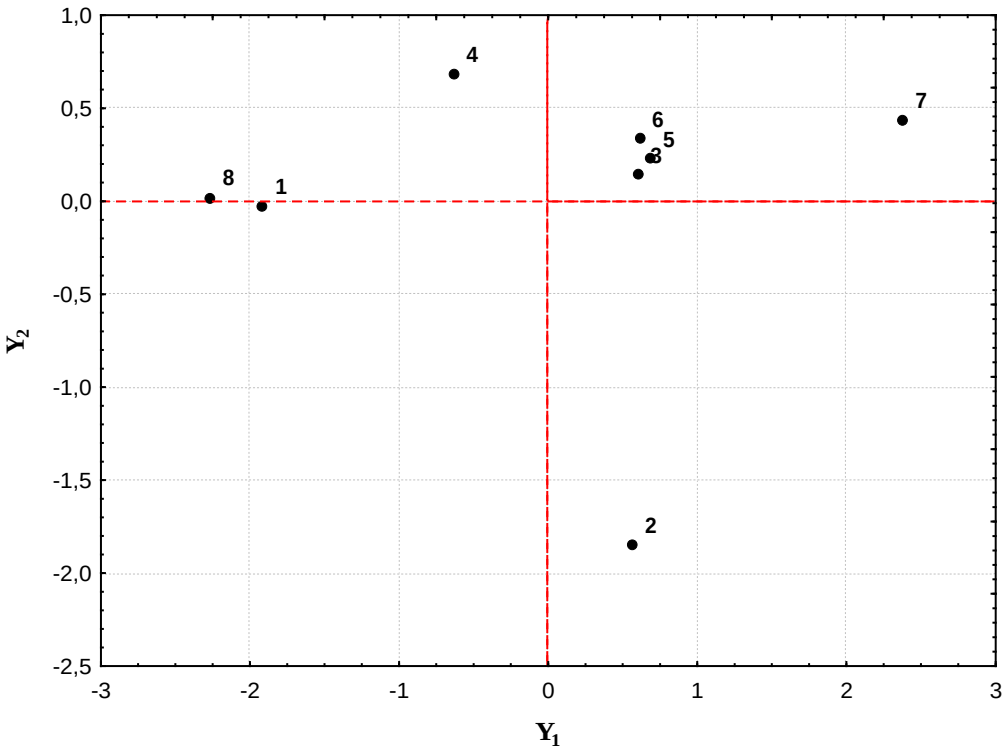

Figura 3.6 Plano fatorial representando os pontos-linha.

Pontos próximos na Figura 3.6, indicam indivíduos com características semelhantes nas variáveis consideradas. Como exemplo pode-se citar os casos 5 e 6.

Pode-se padronizar a variabilidade das CP's. Esta padronização gera o que na análise fatorial são chamados de escores fatoriais (*factor scores*), definido na matriz **F**, de dimensão n x k, por:

$$\mathbf{F}_j = \mathbf{Y}_j \frac{1}{\sqrt{\lambda_j}} \quad (j = 1, 2, ..., k) \tag{3.12}$$

onde j representa a j-ésima coluna da matriz de escores fatoriais.

Aplicando (3.12), tem-se: $\mathbf{F} = \begin{bmatrix} -1{,}254 & -0{,}02 & -0{,}278 \\ 0{,}367 & -2{,}366 & -0{,}393 \\ 0{,}391 & 0{,}189 & -0{,}347 \\ -0{,}416 & 0{,}879 & -1{,}764 \\ 0{,}446 & 0{,}296 & 1{,}097 \\ 0{,}398 & 0{,}434 & 1{,}236 \\ 1{,}551 & 0{,}561 & -0{,}368 \\ -1{,}484 & 0{,}031 & 0{,}818 \end{bmatrix}$

3.9 Plano fatorial para representar as variáveis (pontos-coluna)

Assim como os pontos-linha (observações) podem ser representados num plano fatorial, chamado de plano fatorial das observações, também pode-se representar os pontos-coluna (variáveis) num outro plano, chamado de plano fatorial das variáveis.

As coordenadas para representar as variáveis serão representadas por uma matriz **L**, de dimensão m x k, chamada de matriz das cargas fatoriais (*factor loadings*) cujas colunas são dadas por:

$$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j \quad (j = 1, 2, ..., k). \tag{3.13}$$

Para o exemplo 3.1, aplicando (3.13), a matriz **L**, contendo as coordenadas para variáveis, em três dimensões, será:

$$\mathbf{L} = \begin{bmatrix} 0{,}94756 & -0{,}27872 & 0{,}156334 \\ 0{,}95607 & -0{,}24581 & 0{,}159727 \\ 0{,}72751 & 0{,}68607 & 0{,}006287 \end{bmatrix}$$

em que a primeira linha é a coordenada da variável peso, a segunda linha, a variável altura e a terceira linha a coordenada da variável idade, todas padronizadas.

Cada elemento l_{ij} de **L**, representa a carga fatorial da i-ésima variável na j-ésima componente, tendo que $l_{ij} = corr\left(Z_i, Y_j\right)$.

Considerando-se apenas duas dimensões, tomam-se apenas as duas primeiras colunas de **L**, e assim tem-se:

$$\mathbf{L} = \begin{bmatrix} 0{,}94756 & -0{,}27872 \\ 0{,}95607 & -0{,}24581 \\ 0{,}72751 & 0{,}68607 \end{bmatrix}$$

e sua representação é dada na Figura 3.7. (O Eixo 1 refere-se a primeira coluna de **L** e Eixo 2 à segunda coluna de **L**).

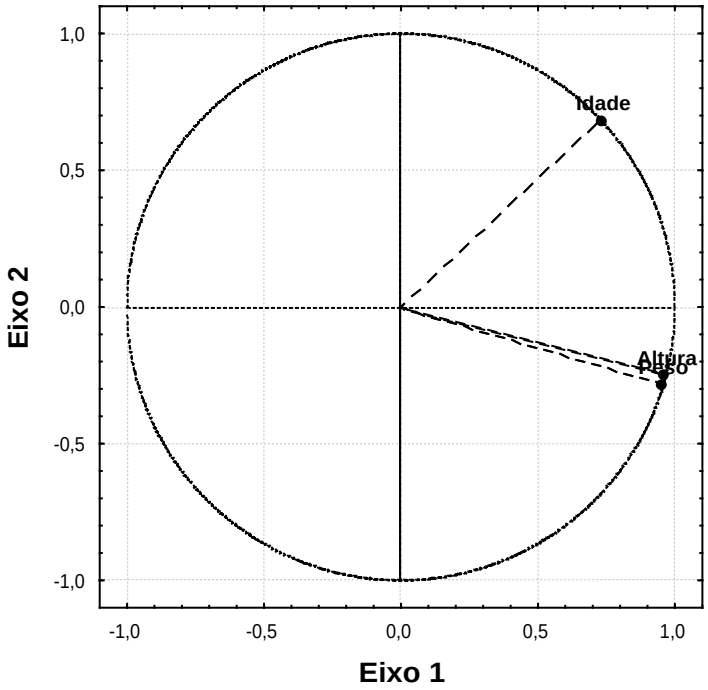

Figura 3.7 : Projeção das variáveis no plano fatorial

Este plano fatorial para representação das variáveis é também conhecido como círculo de correlações, sendo que todos os pontos-variáveis estão contidos em um círculo de raio unitário.

3.10 Sobreposição dos planos fatoriais

Os planos de projeção das observações e das variáveis podem ser representados simultaneamente, segundo BANET e MORINEAU (1999, p.76). Os pontos-coluna (variáveis) são representados por vetores e os pontos-linha (observações) como pontos. Para o exemplo 3.1, tem-se uma representação na Figura 3.8, utilizando o *software* LHSTAT.

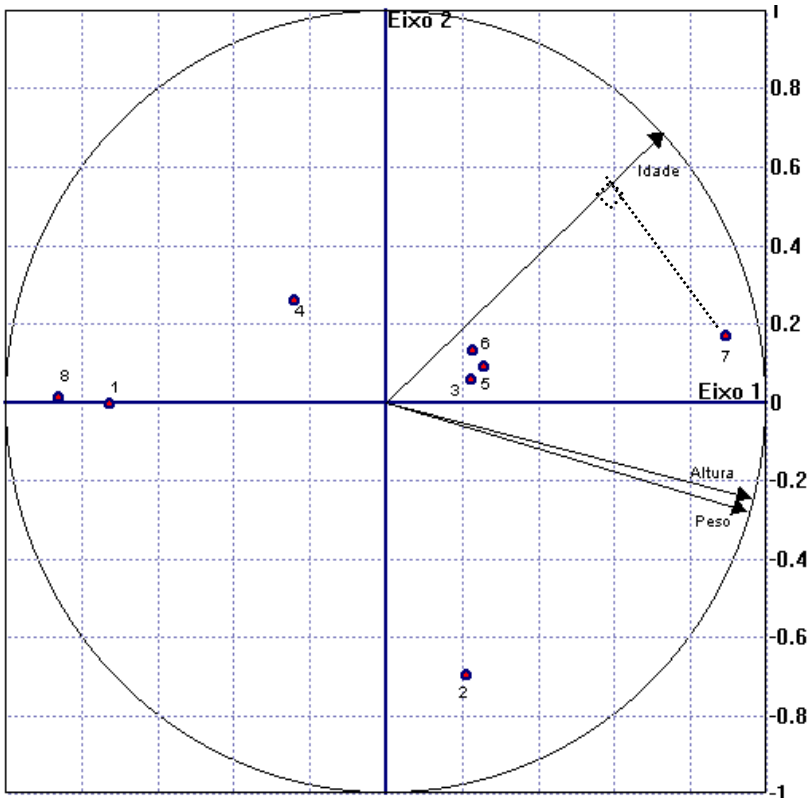

Figura 3.8: Projeção simultânea dos pontos-linha e pontos-coluna

A vantagem da utilização dos planos sobrepostos está na análise das observações com as variáveis conjuntamente.

Pontos próximos à origem do plano representam indivíduos que estão na média do grupo para as variáveis consideradas. Observando a Figura 3.8, os indivíduos 3, 5 e 6

são os que mais se aproximam da origem, portanto são indivíduos com valores mais próximos da média do grupo, para as variáveis consideradas.

Pontos cuja projeção perpendicular sobre os vetores das variáveis recaem próximos à origem, indicam indivíduos que estão na média do grupo naquela variável observada. Como exemplo, pode-se citar o indivíduo 4, considerando a variável idade. Pontos cuja projeção perpendicular sobre os vetores recai acima ou abaixo da origem, indicam indivíduos que estão acima ou abaixo da média, respectivamente, para a variável considerada. Por exemplo, o indivíduo 7 é o que possui mais idade, e sua projeção perpendicular deste sobre o vetor idade recai acima da origem, conforme pode ser visto na Figura 3.8.

3. 11 Correlações das variáveis

O gráfico das variáveis padronizadas permite verificar diretamente o grau de correlação de acordo com o ângulo formado entre elas. O coeficiente de correlação entre duas variáveis, é o cosseno do ângulo formado pelos vetores correspondentes (BANET e MORINEAU,1999, p.29). Assim, ângulo perto de 0 grau: forte correlação positiva; ângulo perto de 180 graus: forte correlação negativa; ângulo perto de 90 graus: correlação nula.

Observando a Figura 3.8, tem-se que os vetores Altura e Peso formam um ângulo próximo de 0 grau, o que indica uma forte correlação positiva entre essas variáveis.

3. 12 Exemplo

Para exemplificar a ACP, usar-se-á um pequeno arquivo de dados, apresentados na Tabela 3.4.

Tabela 3.4 Indicadores demográficos e econômicos - Países do Mundo - 2004

Páis	Pop (milhões hab.)	Densidade (hab/km²)	cresc. Dem (% anual)	Expectativa Vida (anos)	Mort Infantil (%)	Analfabetismo (%)	IDH	PIB (US$ milhões)	Renda Per Capita (US$)	Força Trab (Milhões)	Exportação (US$ milhões)	Importação (US$ Milhões)
Africa do Sul	45,00	36,79	0,59	47,9	47,9	14,8	0,684	113.274	2.820	17,2	29.284	28.405
Alemanha	82,50	231,27	0,07	78,2	4,5	3	0,921	1.846.069	23.560	41	570.791	492.825
Argentina	38,40	13,81	1,17	74,15	20	3,2	0,849	268.638	6.940	15,4	26.655	20.311
Austrália	19,70	2,56	0,96	79,2	5,5	3	0,939	368.726	19.900	9,9	63.387	63.886
Austria	8,10	96,59	0,05	78,45	4,7	3	0,929	188.546	23.940	3,8	70.327	74.428
Bolívia	8,80	8,01	1,89	63,9	55,6	14,6	0,672	7.969	950	3,5	1.285	1.724
Brasil	178,50	20,96	1,24	68,3	38,4	13,1	0,777	502.509	3.070	80,7	58.223	58.265
Camarões	16,00	33,65	1,83	46,25	88,1	28,7	0,499	8.501	580	6,2	1.749	1.852
Canadá	31,50	3,16	0,77	79,3	5,3	3	0,937	694.475	21.930	16,7	259.858	227.165
China	1.304,20	136,24	0,73	71,1	36,6	14,8	0,721	1.159.031	890	763,2	266.155	243.613
Cuba	11,30	101,87	0,27	76,75	7,3	3,3	0,806	25.900	1.000	5,6	1.708	4.930
Dinamarca	5,40	125,31	0,24	76,65	5	2	0,93	161.542	30.600	2,9	51.873	45.398
Espanha	41,10	81,23	0,21	79,35	5,1	2,4	0,918	581.823	14.300	18,2	109.681	142.740
EUA	294,00	31,37	1,03	77,1	6,7	2	0,937	10.065.265	34.280	146,7	730.803	1.180.154
Etiópia	70,70	62,56	2,46	45,45	100,4	60,9	0,359	6.233	100	28,3	420	1.040
França	60,10	110,49	0,47	79	5	3	0,925	1.309.807	22.730	26,8	321.843	325.752
India	1.065,50	324,08	1,51	63,9	64,5	42,8	0,59	477.342	460	460,5	43.611	49.618
Itália	57,40	190,51	-0,1	78,7	5,4	1,6	0,916	1.088.754	19.390	25,8	241.134	232.910
Japão	127,70	342,53	0,14	81,5	3,2	2	0,932	4.141.431	35.610	68,2	403.496	349.089
México	103,50	52,47	1,45	73,4	28,2	8,8	0,8	617.820	5.530	41,3	158.547	176.162
Mongólia	2,60	1,66	1,29	63,9	58,2	1,6	0,661	1.049	400	1,2	250	461
Nigéria	124,00	134,23	2,53	51,45	78,8	36	0,463	41.373	290	51,6	19.150	11.150
Nova Zelândia	3,90	14,42	0,77	78,25	5,8	3	0,917	50.425	13.250	1,9	13.726	13.347
Panamá	3,10	41,05	1,84	74,85	20,6	8,1	0,788	10.171	3.260	1,2	911	2.984
Paraguai	5,90	14,51	2,37	70,85	37	6,7	0,751	7.206	1.350	2,1	989	2.145
Senegal	10,10	51,34	2,39	52,95	60,7	62,6	0,43	4.645	490	4,4	1.080	1.510
Uruguai	3,40	19,29	0,72	75,25	13,1	2,4	0,834	18.666	5.710	1,5	2.060	3.061
Vietnã	81,40	246,99	1,35	69,25	33,6	7,5	0,688	32.723	410	41,1	15.093	15.550

Fonte : Almanaque Abril - 2004

As 12 variáveis são indicadores demográficos e econômicos de 28 países do mundo.

Os indicadores demográficos citados na Tabela 3.4 são:

- **População**: número de habitantes em milhões de pessoas.
- **Densidade**: número de habitantes por km².
- **Crescimento Demográfico**: aumento médio anual do número de indivíduos de uma região, expresso em porcentagem. Resulta do saldo entre os nascimentos e as mortes (crescimento vegetativo) mais o saldo de imigrantes e emigrantes (crescimento migratório).
- **Expectativa de Vida**: estimativa do tempo de vida que a criança terá ao nascer.

- **Mortalidade Infantil**: número de crianças que morrem no primeiro ano de vida entre mil nascidas vivas.
- **Analfabetismo**: proporção de pessoas com 15 anos ou mais que não entendem e/ou não sabem ler nem escrever pequenas frases.
- **IDH** (Índice de Desenvolvimento Humano): mede o bem-estar da população, enfocando três aspectos: vida longa e saudável (expectativa de vida), conhecimento (escolaridade) e padrão de vida decente (PIB per capita – PPC). Sua escala varia de 0 a 1 – quanto mais próximo de 1 melhor a qualidade de vida.

  O IDH, criado no início da década de 90 para o PNUD (Programa das Nações Unidas para o Desenvolvimento) pelo conselheiro especial Mahbub ul Haq, combina três componentes básicos do desenvolvimento humano:

  · a longevidade, que também reflete, entre outras coisas, as condições de saúde da população; medida pela esperança de vida ao nascer.

  · a educação; medida por uma combinação da taxa de alfabetização de adultos e a taxa combinada de matrícula nos níveis de ensino: fundamental, médio e superior.

  · a renda; medida pelo poder de compra da população, baseado no PIB per capita ajustado ao custo de vida local para torná-lo comparável entre países e regiões, através da metodologia conhecida como paridade do poder de compra (PPC).

- **PIB** (Produto Interno Bruto): total de bens e serviços produzidos por um país no período de um ano. Expresso em US$, mede quanto a produção cresceu de um ano para o outro.
- **Renda per Capita**: representa quanto cada habitante receberia em US$ se o valor do produto Nacional Bruto (PNB) de um país fosse distribuído igualmente entre todos.
- **Força de Trabalho**: total de pessoas entre 15 e 64 anos que desempenham atividade remunerada, gerando riquezas para o país.
- **Exportações e Importações**: valor em US$ de todos os bens que um país vende ou compra do resto do mundo.

Como analisar estes dados? Quais conclusões importantes são possíveis obter a respeito das variáveis (indicadores) e das observações (países)? Existem variáveis correlacionadas? Quais os países que se agrupam por possuírem características semelhantes?

Observando o conjunto de dados apresentado na Tabela 3.4, não são evidentes as relações entre as variáveis, entre os casos, e entre os casos e as variáveis. Da maneira como os dados são apresentados, não é possível obter grandes informações. Então, uma tentativa seria recorrer a técnicas que permitem reduzir a dimensionalidade dos dados, conservando um alto percentual de explicação, permitindo usar soluções gráficas ou analíticas com os dados reduzidos em uma menor dimensão.

Para analisar os dados da Tabela 3.4 será utilizado o *software STATISTICA 6.0*.

Inicialmente, tem-se um conjunto de dados composto por uma matriz $\mathbf{X}_{n \times m}$, de dimensão n = 28 observações ou países e m = 12 variáveis.

Através do *STATISTICA* é executada a ACP, encontrando-se a matriz de correlações, cujos resultados estão na Tabela 3.5. A expressão para calcular a matriz de correlações foi apresentada na expressão (3.2).

Tabela 3.5 Correlações entre as variáveis

	POP	DENS	CRESC	EXPEC	MORTI	ANALF	IDH	PIB	RENDA	FORCA	EXP	IMP
POP	1,000	0,414	0,000	-0,039	0,172	0,228	-0,153	0,170	-0,160	0,988	0,205	0,186
DENS	0,414	1,000	-0,301	0,149	-0,085	0,074	0,034	0,160	0,242	0,367	0,313	0,163
CRESC	0,000	-0,301	1,000	-0,684	0,790	0,691	-0,797	-0,216	-0,657	-0,029	-0,435	-0,316
EXPEC	-0,039	0,149	-0,684	1,000	-0,938	-0,796	0,926	0,290	0,648	-0,015	0,435	0,365
MORTI	0,172	-0,085	0,790	-0,938	1,000	0,814	-0,962	-0,306	-0,714	0,138	-0,452	-0,388
ANALF	0,228	0,074	0,691	-0,796	0,814	1,000	-0,884	-0,217	-0,506	0,182	-0,316	-0,275
IDH	-0,153	0,034	-0,797	0,926	-0,962	-0,884	1,000	0,340	0,752	-0,126	0,484	0,423
PIB	0,170	0,160	-0,216	0,290	-0,306	-0,217	0,340	1,000	0,611	0,171	0,847	0,950
RENDA	-0,160	0,242	-0,657	0,648	-0,714	-0,506	0,752	0,611	1,000	-0,149	0,708	0,652
FORCA	0,988	0,367	-0,029	-0,015	0,138	0,182	-0,126	0,171	-0,149	1,000	0,224	0,196
EXP	0,205	0,313	-0,435	0,435	-0,452	-0,316	0,484	0,847	0,708	0,224	1,000	0,948
IMP	0,186	0,163	-0,316	0,365	-0,388	-0,275	0,423	0,950	0,652	0,196	0,948	1,000

Analisando a Tabela 3.5 é possível identificar os coeficientes de correlação entre todos os pares de variáveis. Quanto mais próximos de 1, em módulo, mais forte é o relacionamento. Por exemplo, as variáveis IDH e Mortalidade Infantil (MORTI) apresentam, em módulo, uma correlação igual a 0,962. Já a variável População (POP) apresenta baixa correlação com as variáveis, exceto com a variável força.

É possível utilizar coordenadas para a representação das variáveis num círculo de correlações (k = 2 dimensões). Essas coordenadas podem ser obtidas a partir da expressão (3.13) estão apresentadas na Tabela 3.6.

Tabela 3.6 Coordenadas das variávies (Cargas Fatoriais)

	Eixo 1	Eixo 2
POP	0,036	-0,840
DEN	-0,217	-0,504
CRESC	0,796	-0,122
EXPEC	-0,863	0,206
MORTI	0,901	-0,297
ANALF	0,775	-0,401
IDH	-0,925	0,288
PIB	-0,622	-0,508
RENDA	-0,867	-0,006
FORA	0,009	-0,821
EXP	-0,753	-0,496
IMP	-0,699	-0,503

Através da Tabela 3.6 é possível construir o círculo de correlações, que é apresentado na Figura 3.9.

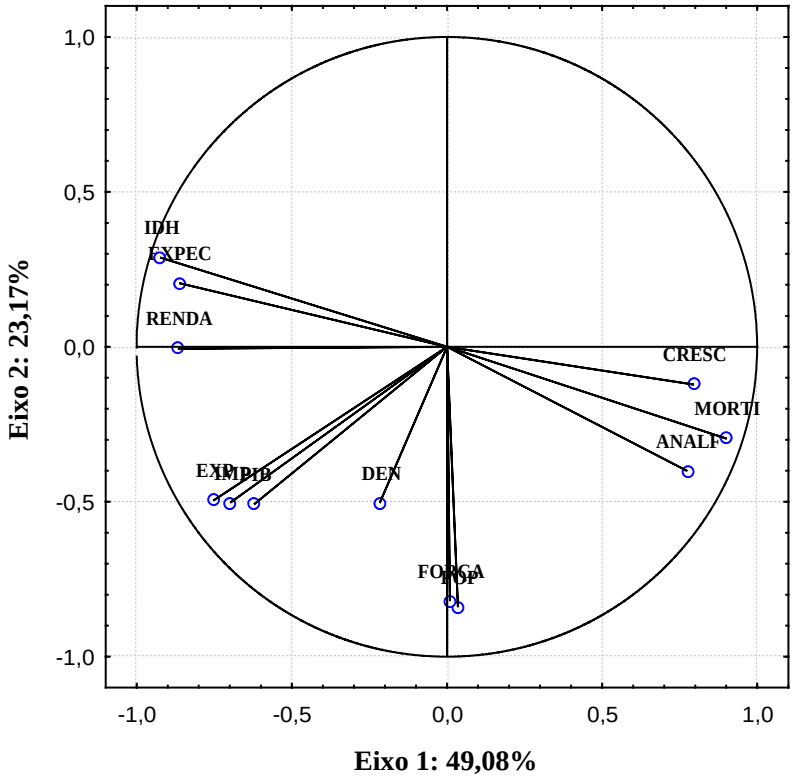

Figura 3.9 Círculo de correlações para as variáveis

O círculo de correlações serve também para estudar a correlação das variáveis, onde cada variável está representada por um vetor, e a proximidade de um vetor do outro, indica um forte relacionamento positivo das variáveis.

Da Figura 3.9, pela proximidade dos vetores é possível perceber a formação de três grupos de variáveis:

Grupo 1: Crescimento Demográfico (CRESC), Mortalidade Infantil (MORTI) e Analfabetismo (ANALF);

Grupo 2: Expectativa de Vida (EXPEC) , IDH e RENDA;

Grupo 3: População (POP), Densidade Demográfica (DENS), Exportação (EXP), Importação (IMP) , PIB e Força de Trabalho (FORÇA).

Em cada grupo as variáveis estão fortemente correlacionadas. As variáveis do grupo 2, também estão fortemente correlacionadas negativamente em relação ao grupo 1. Isto significa que quando um país tem alto IDH e Expectativa de Vida, terá baixo valor para Crescimento Demográfico, Mortalidade Infantil e Analfabetismo.

O conjunto de variáveis do grupo 3 praticamente não se correlaciona com as variáveis dos grupos 1 e 2.

A partir da matriz de correlação da Tabela 3.5 são calculados os autovalores, através da expressão (3.3) que são apresentados na Tabela 3.7.

Tabela 3.7 Autovalores e Inércias

Número do eixo principal	Autovalor	% Inércia total	% Inércia Acumulada
1	5,89	49,08	49,08
2	2,78	23,17	72,25
3	1,60	13,37	85,62
4	0,85	7,05	92,67
5	0,31	2,61	95,28
6	0,26	2,19	97,47
7	0,14	1,13	98,60
8	0,11	0,90	99,50
9	0,03	0,24	99,74
10	0,02	0,16	99,90
11	0,01	0,08	99,98
12	0,00	0,02	100,00
Total	12	100,00	

Com base na Tabela 3.7 e no critério da variância explicada (seção 3.7.3) para a seleção do número de componentes principais, pode-se concluir que as 12 variáveis da Tabela 3.4 podem ser reduzidas a duas componentes, explicando 72,25% da variação dos dados originais.

Também pode ser interessante analisar os países com características semelhantes segundo as variáveis analisadas. Isto pode ser feito através da representação gráfica das novas coordenadas, Figura 3.10, adotando k = 2 dimensões.

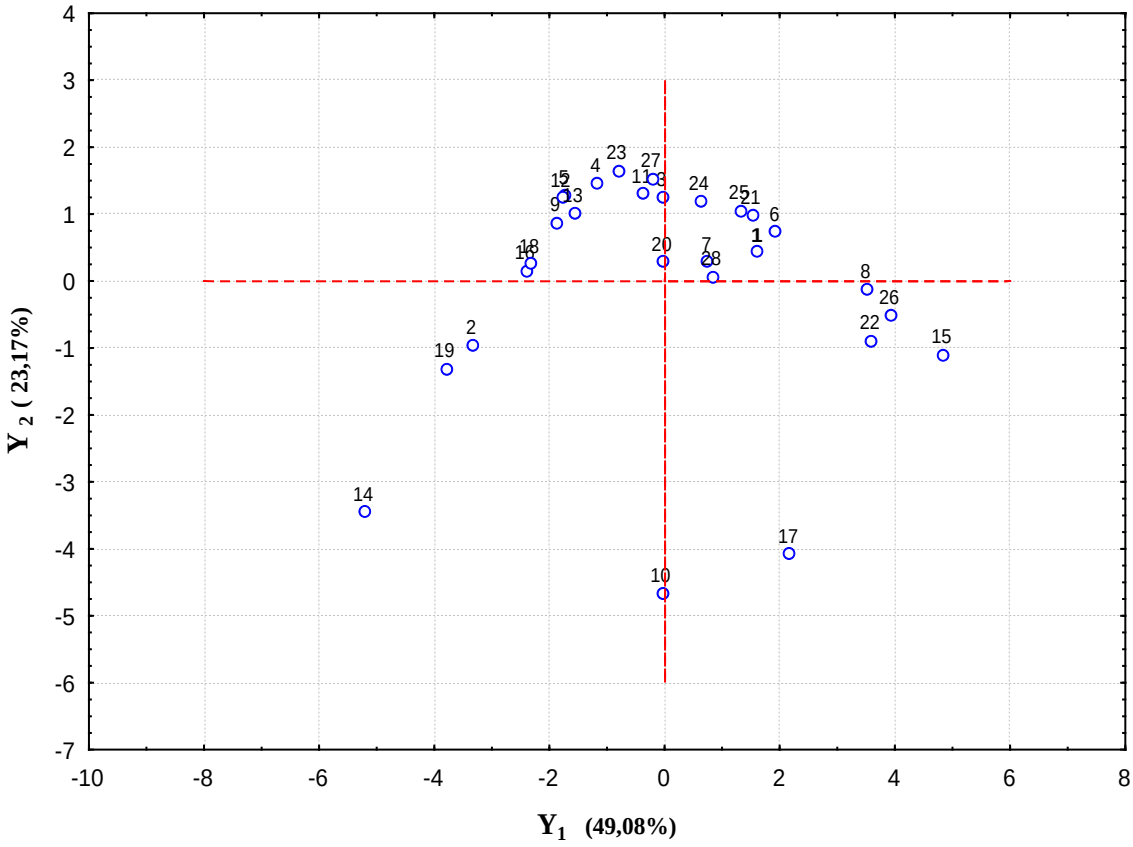

Figura 3.10 Plano fatorial para representação das observações

Os países com características parecidas nas variáveis consideradas, possuem coordenadas também parecidas e no gráfico da Figura 3.10 estão localizados próximos..

Como exemplo pode-se citar os países de número 2 e 19, respectivamente, Alemanha e Japão.

As coordenadas de cada país são dadas na Tabela 3.8 e foram obtidas segundo a expressão (3.11), apresentada na seção 3.8 com maiores detalhes.

Tabela 3.8 Coordenadas dos Casos

Páis	Y_1	Y_2
1 - Africa do Sul	1,605	0,447
2 - Alemanha	-3,345	-0,955
3 - Argentina	-0,036	1,257
4 - Austrália	-1,181	1,464
5 - Austria	-1,731	1,280
6 - Bolívia	1,925	0,733
7 - Brasil	0,747	0,281
8 - Camarões	3,505	-0,125
9 - Canadá	-1,883	0,875
10 - China	-0,038	-4,660
11 - Cuba	-0,366	1,298
12 - Dinamarca	-1,773	1,240
13 - Espanha	-1,562	1,016
14 - EUA	-5,203	-3,448
15 - Etiópia	4,840	-1,096
16 - França	-2,394	0,160
17 - India	2,178	-4,078
18 - Itália	-2,320	0,252
19 - Japão	-3,795	-1,308
20 - México	-0,018	0,283
21 - Mongólia	1,524	0,982
22 - Nigéria	3,599	-0,886
23 - Nova Zelândia	-0,803	1,653
24 - Panamá	0,630	1,190
25 - Paraguai	1,325	1,057
26 - Senegal	3,932	-0,509
27 - Uruguai	-0,202	1,530
28 - Vietnã	0,842	0,066

As colunas da Tabela 3.8, Y_1, Y_2, representam as coordenadas obtidas pelas componentes principais, substituindo as 12 variáveis originais da Tabela 3.4.

Esta técnica pode ser útil para o pré-processamento em mineração de dados pelas seguintes razões:

- reduz a dimensionalidade dos dados, podendo diminuir o tempo computacional para a aplicação de outras técnicas (análise de agrupamentos, classificação, redes neurais, entre outras)
- gera novas variáveis, $Y_1, Y_2, ..., Y_k$, não correlacionadas, acelerando o tempo de certos algoritmos, como o *backpropagation*, utilizado em redes neurais.

As desvantagens da técnica são:

- Ao utilizar as primeiras componentes tem-se perda de informação.
- As CP´s não têm um significado claro como as variáveis originais.

4. ACP COM ESCALONAMENTO ÓTIMO

O objetivo neste capítulo é fazer a revisão da literatura sobre o escalonamento ótimo e sua utilização na análise de componentes principais.

A análise de componentes principais é uma técnica estatística multivariada utilizada para reduzir a dimensionalidade de um conjunto de dados, podendo ser aplicada quando as variáveis são intervalares. Porém, em mineração de dados é comum encontrar variáveis medidas em escalas ordinal e nominal. Por exemplo, quando se deseja avaliar a satisfação dos clientes em relação a um determinando produto. As respostas para as perguntas elaboradas para preferência geralmente constituem uma escala ordinal. Para ilustrar, considere uma pesquisa em que está sendo verificada a satisfação dos clientes que adquiriram um carro zero km, em que é feita a seguinte pergunta aos proprietários: Na sua opinião, como você classificaria o conforto do seu automóvel? As opções de respostas são:

( ) Ruim ( ) Regular ( ) Bom ( ) Ótimo

Essa pergunta mede algo subjetivo a cada pessoa, cujas respostas estão apresentadas numa escala ordinal de satisfação: Ruim, Regular, Bom e Ótimo.

É usual, quando as variáveis são medidas em escalas do tipo ordinal ou nominal , atribuir códigos numéricos para as categorias de cada variável. A quantificação pode ser usada como uma opção para se usar as técnicas tradicionais que, normalmente, supõem variáveis quantitativas.

Porém, quais os valores a serem atribuídos na quantificação: Ruim: 1, Regular: 2, Bom: 3 e Ótimo: 3? ou Ruim: 4, Regular: 6 Bom: 7 e Ótimo: 9? Outra seqüência qualquer? Tanto faz? Apesar da arbitrariedade, ao atribuir os valores 1, 2, 3 e 4 para a quantificação está se considerando uma eqüidistância entre os intervalos, por exemplo Ruim-Regular com mesma distância de Bom-Ótimo, o que não funciona bem assim (BELL, 2004). Além disso, segundo YOUNG (1981, p.358) existem valores, chamados de “valores ótimos” que maximizam a eficiência do modelo. No caso em estudo, o

modelo de interesse é o de componentes principais, e o objetivo é de obter uma redução de dimensionalidade, preservando a maior variabilidade.

Encontrar tais valores para as categorias de cada variável, chamados de "valores ótimos", é o processo conhecido como escalonamento ótimo. Após a obtenção desses valores otimamente escalonados, aplica-se a análise de componentes principais, como etapa de pré-processamento, cujo objetivo continua sendo de obter k novas coordenadas para as m variáveis observadas, sendo $k < m$, obtendo assim uma redução na dimensionalidade, porém maximizando a inércia ou variância explicada.

4.1 Escalonamento ótimo

YOUNG (1981,p.358) define o escalonamento ótimo como:

> "uma técnica de análise de dados que atribui valores numéricos para as categorias das observações de forma a maximizar a relação entre as observações e o modelo de análise, considerando as características de medida dos dados ".

A definição é genérica, não especificando o modelo de análise (componentes principais, regressão,...) nem a escala de medida das variáveis (intervalar, ordinal ou nominal). Trabalhando com esta definição, Forrest Young, Jan de Leeuw e Yoshio Takane, desenvolveram um grupo de programas/algoritmos para quantificar dados qualitativos, denominando de programas ALSOS (*Alternating Least Squares Optimal Scaling* – Mínimos Quadrados Alternados com Escalonamento Ótimo). Estes programas têm como objetivo determinar valores para as categorias, que maximizam a qualidade do modelo ajustado. Neste contexto, estes valores são chamados de valores ótimos.

Segundo YOUNG (1981, p.359), dependendo do objetivo da análise utiliza-se um dado programa/algoritmo. Por exemplo, se o objetivo é realizar uma análise de componentes principais, o programa/algoritmo desenvolvido foi denominado de PRINCALS & PRINCIPALS.

O escalonamento ótimo tem como objetivo atribuir valores numéricos para as categorias de cada variável, permitindo, deste modo, padronizar procedimentos para serem usados, a fim de obter uma solução através das novas variáveis quantificadas. Os

valores otimamente escalonados são atribuídos para as categorias de cada variável, baseando-se no critério de otimização do algoritmo em uso. Ao contrário das variáveis originais, que podem ser nominais ou ordinais, esses valores escalonados têm propriedades métricas (MEULMANN e HEISER, 2001).

De acordo com estes autores, apesar de existirem modelos para analisar especificamente dados categorizados, eles não geram bons resultados quando o conjunto de dados a ser analisado tem as seguintes características:

- poucas observações;
- muitas variáveis;
- muitos valores por variável.

Na maioria dos algoritmos que tratam variáveis categorizadas, os valores ótimos para cada variável escalonada são obtidos através de um método iterativo chamado mínimos quadrados alternados, no qual os valores correntes são usados para encontrar uma solução, que serão substituídos pela solução encontrada (MEULMAN e HEISER, 2001,p.1). Os valores atualizados são usados para encontrar uma nova solução, a qual é usada para atualizar os valores, e assim por diante, até que algum critério de convergência seja alcançado, sinalizando que o processo deve parar. A Figura 4.1 ilustra o funcionamento do escalonamento ótimo nos programas ALSOS (YOUNG, 1981, p.360).

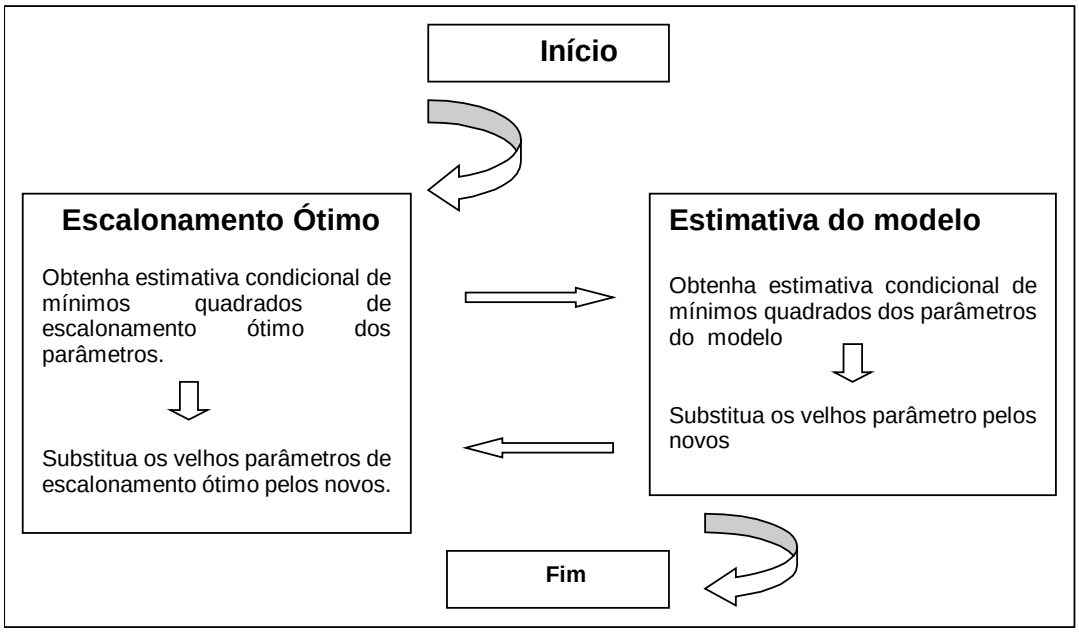

Figura 4.1 Fluxograma dos programas ALSOS

Fonte: YOUNG (1981)

Encontra-se na versão 11.0 do *software* SPSS a implementação destes algoritmos, realizada por um grupo de pesquisadores da Universidade de Leiden, *Data Theory Group* – DTG, membros do departamento de Educação e Psicologia da Faculdade de Ciências Sociais e do Comportamento desta Universidade (DTG, 2005). Este grupo apoiou-se em estudos de escalonamento multidimensional e análise quantitativa para variáveis qualitativas, feitos por Forrest YOUNG (2005), Yoshio TAKANE (2005) e Jan de LEEUW (2005).

Quando as variáveis são todas quantitativas e são escalonadas no nível numérico, os resultados do escalonamento ótimo são equivalentes a ACP tradicional, utilizada para variáveis quantitativas. E, quando as variáveis são todas qualitativas, e as variáveis são escalonadas no nível nominal, o método equivale a análise de correspondência múltipla (MEULMANN e HEISER, 2001, p.9).

4.2 O algoritmo

A análise de componentes principais para dados categóricos, denominada de CATPCA – *Categorical Principal Components Analysis*, encontrada no SPSS, versão 11, executa simultaneamente duas tarefas: quantificar variáveis categóricas e reduzir a dimensionalidade dos dados. Esta quantificação é realizada seguindo o escalonamento ótimo, que resultará em componentes principais ótimos para as variáveis transformadas. As variáveis podem ter medidas em diferentes níveis: nominal, ordinal ou intervalar (MEULMANN e HEISER,2001, p.27).

O objetivo de se utilizar o escalonamento é encontrar os valores ideais para as categorias de cada variável, a fim de realizar um melhor ajuste entre o relacionamento das variáveis, entre os casos, ou entre casos e variáveis, uma vez que a simples atribuição de valores numéricos às categorias é apenas um número atribuído pelo pesquisador, o que pode ficar difícil de julgar as similaridades e dissimilaridades entre os casos, entre os objetos ou casos e objetos.

Para realizar a busca dos valores ideais para as categorias de cada variável, o algoritmo utiliza-se de um processo iterativo sobre uma matriz de dados otimamente escalonados, alternando com modelo de mínimos quadrados (regressão, componentes principais, etc.), até atingir um resultado satisfatório. Neste trabalho, o modelo de interesse é o de componentes principais. O controle do processo se dá até que a convergência do método seja atingida (YOUNG *et al.*,1978, p.280).

Conforme YOUNG *et al.* (1978, p.279-280), o método de análise de componentes principais com escalonamento ótimo inicia com uma matriz $\mathbf{Z}_{nxm}$, sendo n o número de observações padronizadas relativas a m variáveis, que pode ser aproximada (estimada) por uma matriz $\mathbf{Z}^{"}$, definida por :

$$\mathbf{Z}^{"} = \mathbf{F}\mathbf{L}^{\mathbf{t}} \qquad (4.1)$$

onde a matriz **F**, de dimensão n x k, contém os *n* escores relativos aos *k* componentes principais; e a matriz **L**, de dimensão m x k, contém as cargas fatoriais das *m* variáveis em termos dos *k* componentes.

Para que o problema tenha solução única, supõe-se que **F** e **L** satisfaçam as seguintes condições:

1) $\frac{1}{n-1}\mathbf{F}^t\mathbf{F} = \mathbf{I}$ (matriz identidade)

e

2) $\mathbf{L}^t\mathbf{L} = \mathbf{D_k}$ (matriz diagonal)

Para YOUNG *et al*. (1978, p.279-280), as matrizes **F** e **L** podem ser obtidas minimizando a expressão:

$$\theta^* = tr\left[\left(\mathbf{Z}^* - \mathbf{Z}^{''}\right)^t\left(\mathbf{Z}^* - \mathbf{Z}^{''}\right)\right] \qquad (4.2)$$

para um número *k* prescrito de componentes principais.

onde $\mathbf{Z}^*$ é definido como uma matriz n x m de observações otimamente escalonadas, cujas colunas são centradas e normalizadas:

$$\mathbf{Z}^{*t}.\mathbf{J}_n = \mathbf{O}_m \text{ e } diag\left[\frac{\mathbf{z}^{*t}\mathbf{z}^*}{n}\right] = \mathbf{J}_m \ , \qquad (4.3)$$

sendo $\mathbf{J}_n$ vetor de ordem n de uns, $\mathbf{J}_m$ vetor de ordem m de uns e $\mathbf{O}_m$ vetor de ordem m de zeros.

Alternativamente, as matrizes **F** e **L** podem ser obtidas conforme visto no capítulo 3, segundo as expressões (3.12) e (3.13) respectivamente.

A expressão 4.2 é a versão multivariada da soma dos quadrados dos erros, muito usada em análise de regressão e também em componentes principais.

A aplicação de (4.3) corresponde a subtrair de cada observação a média aritmética e dividir pelo desvio padrão da respectiva coluna (padronização).

O processo de minimização dado por (4.2) é conhecido como método de mínimos quadrados, análogo ao processo apresentado na seção 3.4 do capítulo 3, em que se busca minimizar a inércia ou variância não explicada.

Um dos algoritmos desenvolvidos por Young, Takane e De Leeuw é o chamado PRINCIPALS (*Principal Components Analysis by Alternating Least Squares*). Nesta seção descreve-se o algoritmo e no Apêndice 1 apresenta-se um exemplo numérico. O procedimento PRINCIPALS otimiza θ^* (expressão 4.2) sob a restrição feita em (4.3). É

baseado no princípio de mínimos quadrados alternados (ALS) e consiste de duas fases: uma fase de estimação do modelo (otimização de θ^* com relação aos parâmetros do modelo (elementos da matriz **F** e **L**) e a fase de escalonamento ótimo (otimização de θ^* com relação aos parâmetros de dados $\mathbf{Z}^*$). As duas fases são alternadamente iterativas até que a convergência seja obtida.

A Figura 4.2 mostra o procedimento de otimização pelo método de mínimos quadrados com escalonamento ótimo.

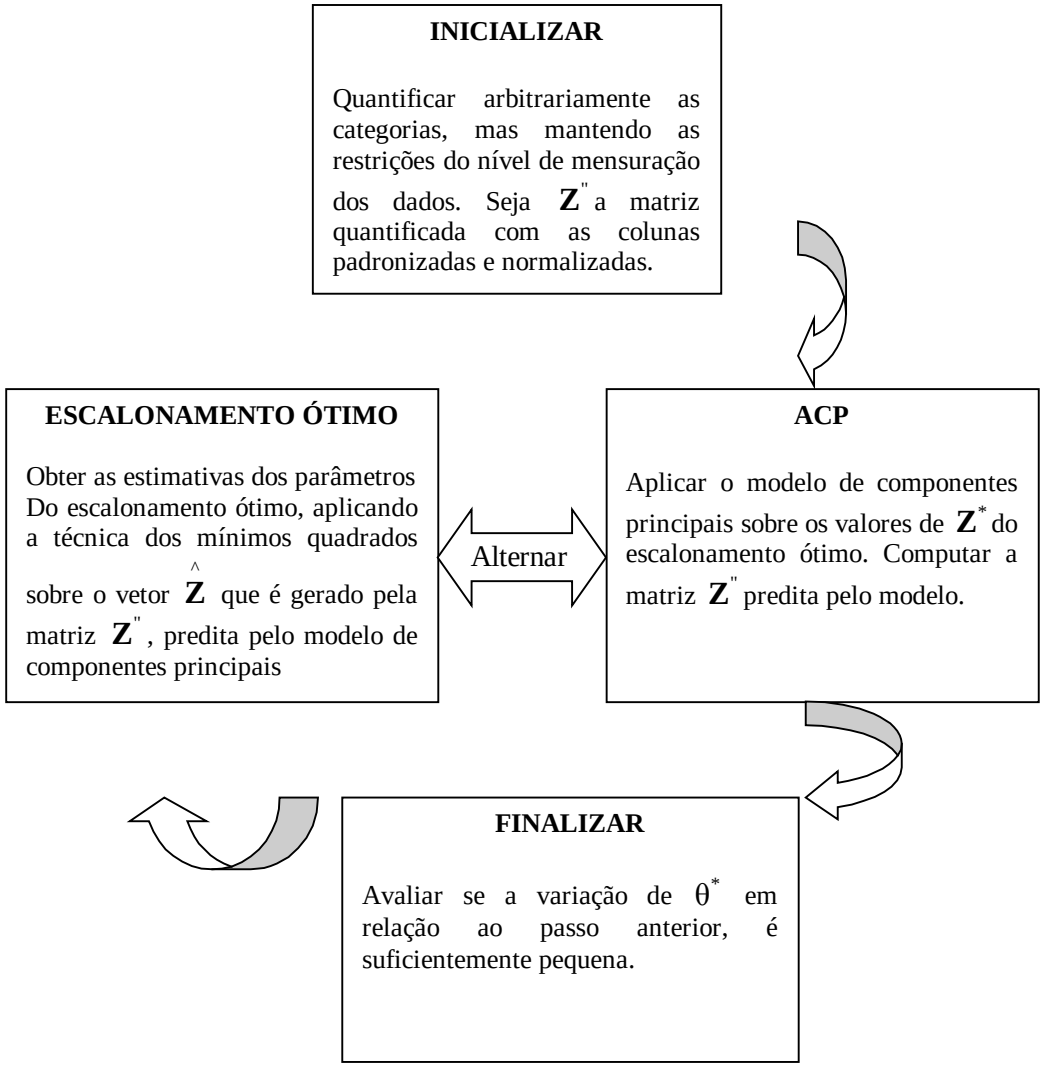

Figura 4.2 Esquema do funcionamento dos mínimos quadrados alternados

Fonte: YOUNG (1981), adaptado por Ivanqui e Barbetta (2002)

Seguindo YOUNG *et al*. (1978), os passos do algoritmo são os seguintes:

Passo 0: Início

Criar a matriz de observações, chamada matriz **X**, onde cada linha de **X** corresponde a uma observação, e, cada coluna a uma variável.

Quando as variáveis da análise não são quantitativas, para iniciar o processo é necessário atribuir valores às categorias das variáveis. Os valores a serem atribuídos são arbitrários, e se a variável for nominal nenhuma ordem precisa ser respeitada. Porém, se a variável for ordinal, esses valores devem respeitar uma ordinalidade.

Passo 1: Padronização e normalização dos dados

Os dados observados **X** são padronizados e normalizados para uma matriz $\mathbf{Z}^*$ e usados como uma matriz inicial. Para variáveis nominais são atribuídos arbitrariamente números às categorias; e para as variáveis ordinais também são atribuídos números às categorias, mas mantendo a restrição da ordenação.

As expressões para a padronização e normalização dos dados podem ser encontradas no capítulo 3, respectivamente (3.1) e (3.5).

Passo 2: Estimativa do modelo

Calcula-se a matriz de correlações de $\mathbf{Z}^*$:

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^* \qquad (4.4)$$

A partir de **R** determina-se o número de componentes *k* que serão retidos para o processo e calcula-se $\mathbf{D_k}$ e $\mathbf{V_k}$, em que :

$\mathbf{D_k}$ é a matriz diagonal formada pelos *k* maiores autovalores de **R**; e

$\mathbf{V_k}$ é a matriz cujas colunas são os autovetores normalizados de $\mathbf{Z}^{*t}\mathbf{Z}^*$, correspondentes aos *k* maiores autovalores.

Daí, pode-se calcular as matrizes **F** e **L** pelas expressões :

$$\mathbf{L}_j = \sqrt{\lambda_j}.\mathbf{v_j} \tag{4.5}$$

em que λ_j corresponde ao j-ésimo autovalor de $\mathbf{R}$ e $\mathbf{v}_j$ corresponde ao j-ésimo autovetor associado a λ_j, com j = 1,2, ...,k.

$$\mathbf{F} = \mathbf{Z}^*\mathbf{L}\mathbf{D}_\mathbf{k}^{-1} \tag{4.6}$$

Este processo é equivalente ao apresentado no capítulo 3, segundo a expressão (3.11).

Passo 3: Avaliação do Modelo

Calcula-se θ^* , através da expressão (4.2). A partir da segunda iteração, o modelo é avaliado, comparando-se θ^* do passo atual com o θ^* do passo anterior. Se a diferença entre os θ^*´s for desprezível, então encerra-se o processo.

Passo 4: Estimação dos dados escalonados

Estimativa dos parâmetros de escalonamento ótimo: Através de **F** e **L** é calculado $\mathbf{Z}^{"}$ através da expressão (4.1). Obtém-se, então, a matriz de escalonamento dos dados, $\mathbf{Z}^*$, tal que θ^* é minimizado para $\mathbf{Z}^{"}$ fixo, dentro das restrições de medida de cada variável.

O escalonamento ótimo pode ser feito para cada variável separadamente e independentemente, assim θ^* é separável com relação ao escalonamento ótimo de dados para cada variável. Isto significa que a expressão (4.2) pode ser reescrita como uma soma de parcelas independentes, uma para cada variável:

$$\theta^* = \sum_{j=1}^{m} \left(\mathbf{z}_j^* - \mathbf{z}_j^{"}\right)^t \left(\mathbf{z}_j^* - \mathbf{z}_j^{"}\right) = \sum_{j=1}^{m} \theta_j^* \tag{4.7}$$

onde $\mathbf{Z}_j^*$ e $\mathbf{Z}_j^{"}$ são o j-ésimo vetor coluna de $\mathbf{Z}^*$ e $\mathbf{Z}^{"}$, respectivamente.

Para o escalonamento ótimo serão considerados os seguintes vetores e matrizes :

- um vetor **B**, de dimensão n.m, contendo os valores ordenados da matriz $\mathbf{Z}^{*}_{n\,x\,m}$,
- um vetor $\hat{\mathbf{Z}}$, de dimensão n.m, contendo os valores ordenados da matriz $\mathbf{Z}^{"}_{n\,x\,m}$, numa ordem que mantenha uma correspondência biunívoca com os elementos do vetor **B**;
- uma matriz **G** binária, denominada matriz indicadora, composta por n.m colunas e tantas linhas quantas forem as categorias $a_1, a_2, a_3, ..., a_r$ observadas na escala. (máximo de r linhas)

  A primeira linha de **G** tem valor 1 somente nas posições em que **B** tenha o menor valor e 0 nas outras posições; a segunda linha de **G** tem valor 1 nas posições em que **B** tenha o segundo maior valor e 0 nas outras posições; e assim por diante.

Obtém-se os valores não normalizados do escalonamento, dados pelo vetor $\mathbf{Z}^{G}$, através do operador de projeção de mínimos quadrados aplicado ao vetor $\hat{\mathbf{Z}}$, definido por:

$$\mathbf{Z}^{G} = \mathbf{G}\left(\mathbf{G}^{t}\mathbf{G}\right)^{-1}\mathbf{G}^{t}\,\hat{\mathbf{Z}}^{t}. \qquad (4.8)$$

Para preservar a ordinalidade, caso os elementos de $\mathbf{Z}^{G}$ não estiverem na ordem representada pelo vetor **B**, deve-se fazer a ordenação de $\mathbf{Z}^{G}$. A matriz de escalonamento $\mathbf{Z}^{*}_{n\,x\,m}$ é composta pelos elementos de $\mathbf{Z}^{G}$, de dimensão m.n organizados por linha.

Os valores gerados através de (4.8) serão os novos elementos da matriz **X**, e retorna-se ao Passo 1, e o processo se repete até que a melhoria no ajuste da iteração anterior para a presente iteração seja desprezível.

Conforme já comentado no Passo 0, quando as variáveis não são quantitativas, atribuí-se arbitrariamente valores as categorias para executar o escalonamento ótimo.

Apesar dos valores serem atribuídos arbitrariamente, as conclusões serão as mesmas, independentemente dos valores atribuídos, influenciando apenas no valor de teta e no número de iterações para que se atinja o critério de convergência. Dependendo dos valores iniciais será maior ou menor o número de iterações para um mesmo critério de convergência ser atingido.

No Anexo 1 é apresentado um exemplo numérico ilustrando os passos do algoritmo, considerando variáveis ordinais e nominais.

5. PRÉ-PROCESSAMENTO PARA A MINERAÇÃO DE DADOS

Neste capítulo será discutido brevemente o pré-processamento em seu aspecto mais amplo, e, também a importância da análise exploratória de dados. Em seguida, será apresentada uma metodologia específica de pré-processamento para a mineração de dados.

5.1 Pré-processamento

Conforme apresentado no capítulo 3, seção 2.4, o pré-processamento em seu sentido mais amplo tem como objetivo melhorar a qualidade dos dados para a mineração, pois é comum encontrar nas bases de dados registros incompletos, valores inconsistentes, valores discrepantes entre outros problemas. Este pré-processamento inclui tarefas como a limpeza, imputação, integração, transformações e redução dos dados. Para auxiliar nestas tarefas do pré-processamento é comum utilizar-se a análise exploratória de dados.

5.2 Análise exploratória

Antes de aplicar qualquer técnica estatística mais avançada, um estudo inicia-se através da análise exploratória dos dados (AED). Esta análise visa conhecer características da base, resumindo-as em tabelas e gráficos apropriados para cada tipo de variável. A exploração dos dados, através da estatística descritiva, permitirá descobrir como os valores estão distribuídos, detectar a existência de valores discrepantes (*outilers)* e identificar registros incompletos ou faltantes (*missing values*).

Tendo o conhecimento da base de dados o pesquisador deverá decidir como tratar os registros com valores faltantes e com valores discrepantes. Alguns procedimentos são apresentados na seção 2.3.1.

Nesta etapa, todas as variáveis devem ser analisadas, levando-se em conta o nível de mensuração da variável. Em seguida, são apresentadas algumas sugestões para realizar a exploração.

Variáveis mensuradas qualitativamente podem ser resumidas utilizando distribuição de freqüências e representações gráficas tais como, gráficos de barras e setores. Um exemplo para variável qualitativa é apresentado na Tabela 5.1 e na Figura 5.1.

Tabela 5.1 Distribuição de Freqüências do Estado Civil dos pesquisados na cidade de Lages - SC

Estado Civil	Número de pessoas	Percentagem
Casado (a)	4.871	50,81
Separado (a)Judicialmente	307	3,20
Divorciado (a)	274	2,86
Viúvo (a)	634	6,61
Solterio (a)	3.500	36,51
TOTAL	9.586	100,00

Fonte: IBGE 2000

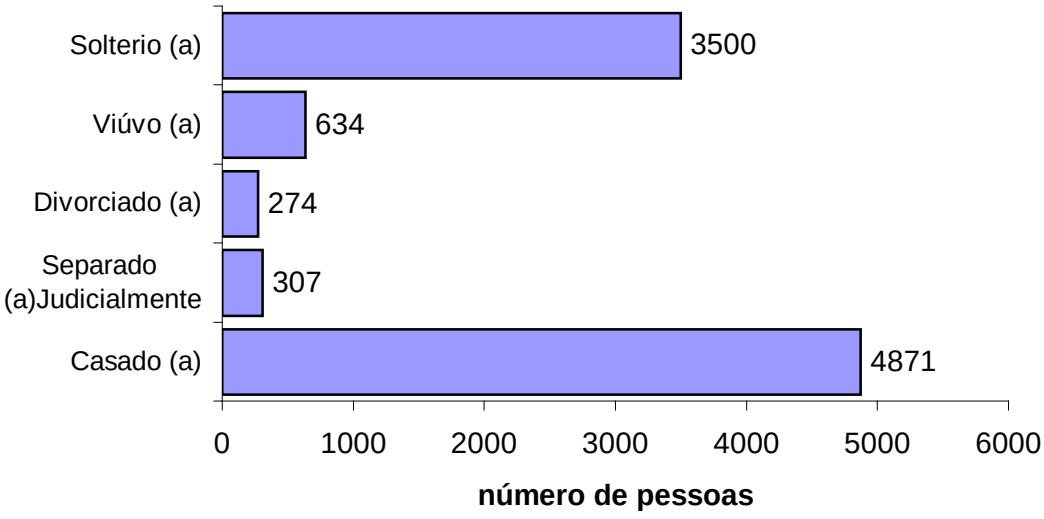

Figura 5.1 Gráfico de barras para a variável Estado Civil

Para as variáveis mensuradas quantitativamente, pode-se utilizar as distribuições de freqüências, recursos gráficos como histograma ou polígono de freqüências. A Tabela

5.2 é um exemplo de distribuição de freqüências para uma variável quantitativa, assim como o histograma da Figura 5.2.

Tabela 5.2 Distribuição de freqüências da Idade dos pesquisados na cidade de Lages - SC

Idade	Número de Pessoas	Percentagem
0 ⊢ 20	875	9,13
20 ⊢ 40	4.708	49,11
40 ⊢ 60	2.850	29,73
60 ⊢ 80	1.046	10,91
80 ⊢ 100	104	1,08
100 ⊢ 120	3	0,03
TOTAL	9.586	100,00

Fonte: IBGE - 2000

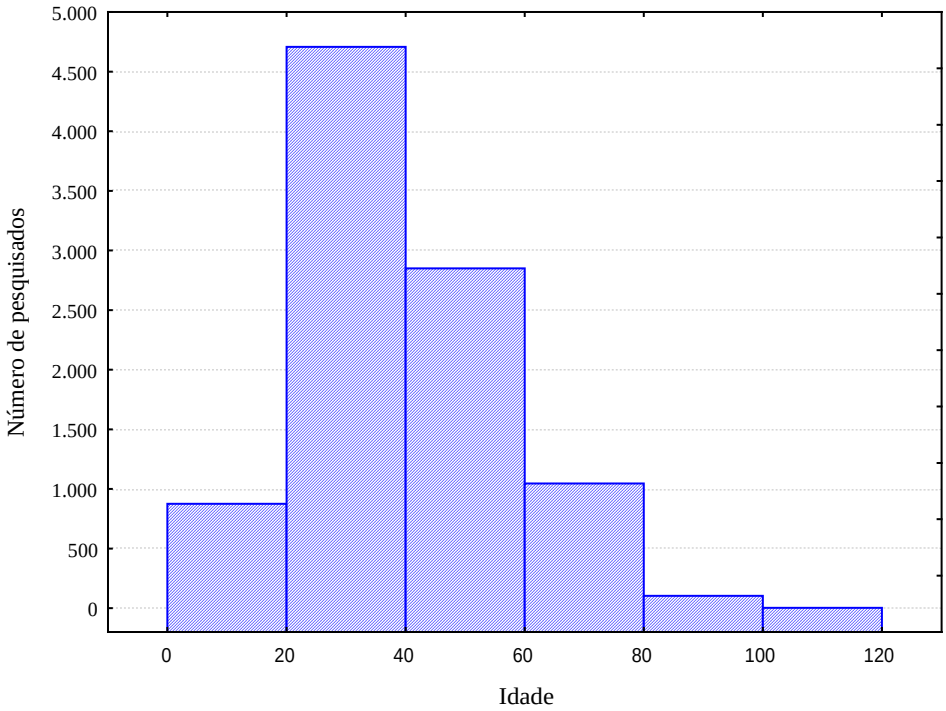

Figura 5.2 Histograma para a variável Idade

As medidas descritivas, tais como média, mediana, moda e o desvio padrão também fornecem um resumo das características em relação à tendência central e da dispersão

dos dados. Algumas medidas resumo em relação à variável Idade (Tabela 5.2), obtidas a partir do *software* Statistica 6.0, são apresentadas na Tabela 5.3.

Tabela 5.3 Medidas resumo da tabela 5.2 em relação a variável Idade

Estatística	Valor
Média	39,66
Mediana	37,00
Moda	18,00
Valor Mínimo	18,00
Valor Máximo	120,00
Desvio Padrão	15,78
Coef. de Variação	40%

Fonte: *Software Statistica* 6.0

Outro recurso gráfico muito utilizado para variáveis quantitativas é o diagrama de caixas, desenho esquemático ou *Box Plot*. Este recurso fornece um excelente resumo visual sobre muitos aspectos importantes da distribuição, tais como dispersão, assimetria e dados discrepantes. Um exemplo de *Box Plot* para a variável Idade é apresentado na Figura 5.3.

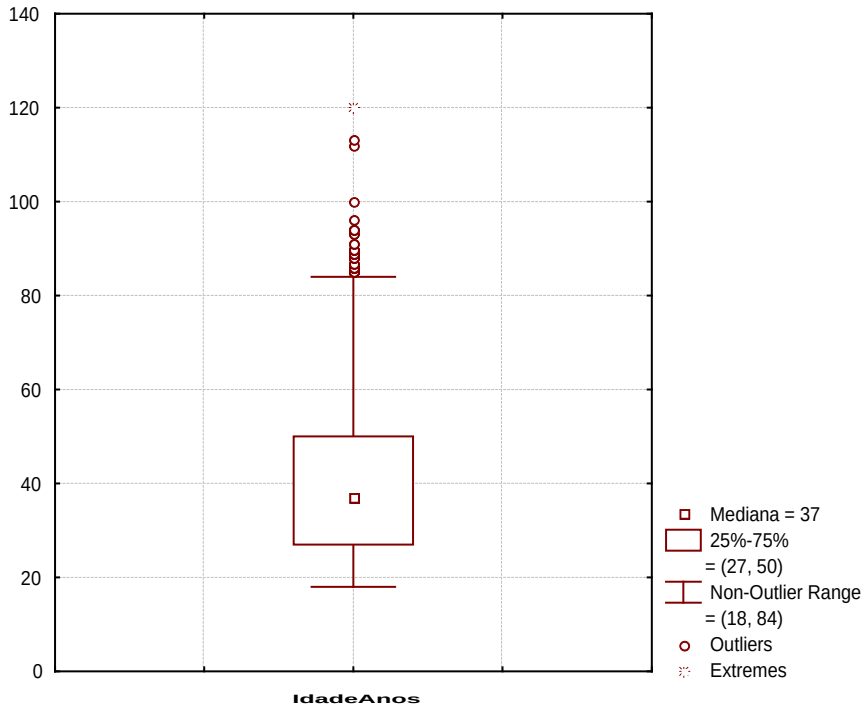

Figura 5.3 *Box Plot* para a variável Idade

Oferecendo informações resumidas, a análise exploratória dá uma visão geral da base de dados. Além disso, a AED serve de auxílio para a verificação da normalidade dos dados. Muitas vezes, em estatística, necessita-se que os dados tenham uma distribuição normal ou, então, aproximadamente normal.

Para investigar a normalidade dos dados pode-se utilizar o histograma ou *box plot*. O histograma de uma distribuição normal é simétrico, tendo formato de sino. Quando a variável e quantitativa discreta, outra maneira é observar a normalidade pode ser pelas medidas de tendência central: Média, Mediana e Moda. Em distribuições perfeitamente normais esses valores coincidem.

Além das opções anteriores, pode-se utilizar o gráfico de probabilidade normal. Este tipo de gráfico pode ser facilmente obtido em *softwares* de estatística. Se os pontos marcados no gráfico estiverem distribuídos em uma linha imaginária, ou próximos de uma linha reta teórica, tem-se que a variável tem distribuição normal ou pelo menos aproximadamente normal. Se os pontos marcados no gráfico parecem se desviar de algum modo dessa linha reta teórica, tem-se que a variável não apresenta uma distribuição normal. As Figuras 5.4 e 5.5 ilustram a idéia.

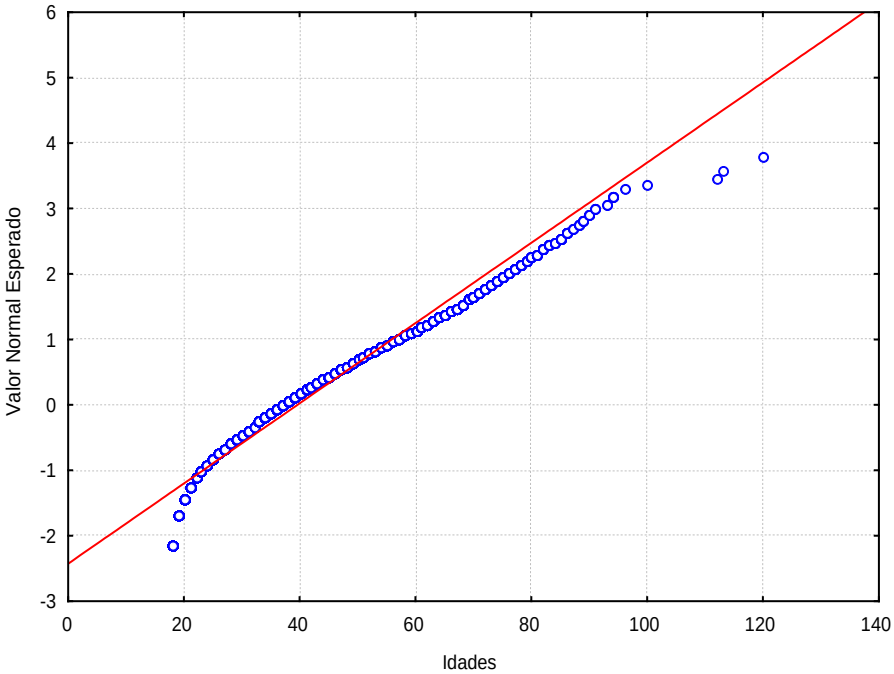

Figura 5.4 Gráfico da probabilidade normal para variável Idade

Observando a Figura 5.4 é possível perceber que os valores se distribuem aproximadamente sobre a linha teórica, assim tem-se que a variável idade apresenta uma distribuição aproximadamente normal.

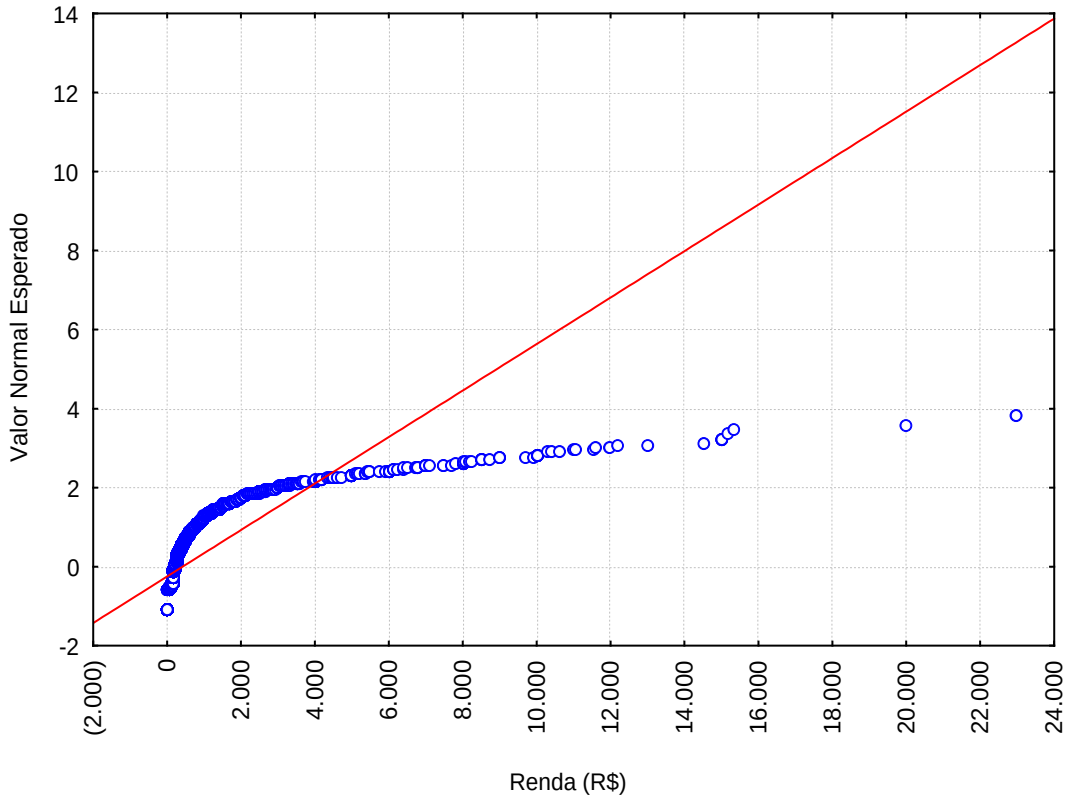

Figura 5.5 Gráfico da probabilidade normal para variável Renda

Já para a variável renda, pode-se observar na Figura 5.5 que valores não se distribuem aproximadamente sobre a linha teórica, assim não se distribuem de maneira normal ou simétrica.

Nestes casos, uma alternativa para tornar a distribuição simétrica pode estar em utilizar uma transformação na variável (seção 2.4.3). Por exemplo, ao invés de usar a variável renda, utilizar o logaritmo desta variável.

Um estudo mais detalhado sobre análise exploratória, tipos de gráficos, cálculo das medidas resumo, propriedades da distribuição normal e o pressuposto de normalidade pode ser encontrado em BARBETTA (2001, p. 69-120); BUSSAB e MORETIN (2003, p. 1–37) e LEVINE *et al*. (2000 p. 219; 240-242).

5.3 Proposta de pré-processamento para a mineração de dados

Nesta seção será apresentada uma metodologia de pré-processamento para a mineração de dados, que permite a entrada de variáveis em diferentes níveis de mensuração, gerando um novo conjunto de variáveis quantitativas. A metodologia também contribui para gerar um conjunto de variáveis, menor que o original, o que pode ser uma solução para reduzir o tempo computacional dos algoritmos de mineração de dados.

A Figura 5.6 ilustra, resumidamente, em duas etapas, a idéia do pré-processamento, quando as variáveis da análise são apenas variáveis quantitativas.

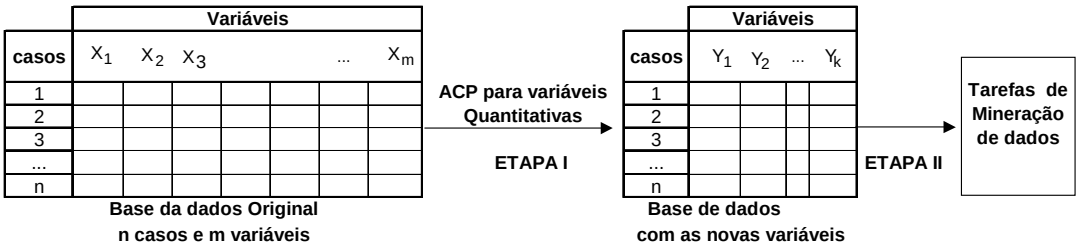

Figura 5.6 Pré-processamento utilizando a ACP

Etapa I: Redução da dimensionalidade através da aplicação da análise de componentes principais (ACP) na base de dados original, com m variáveis $(X_1, X_2, X_3, ..., X_m)$, gerando um novo conjunto de k variáveis $(Y_1, Y_2, ..., Y_k)$, sendo k <m.

Etapa II: Execução de tarefas da mineração de dados, tais como associação, sumarização ou segmentação (análise de agrupamentos). Mais detalhes sobre as tarefas de mineração ver a seção 2.2.

Para aplicar a técnica da ACP é necessário que as variáveis sejam todas mensuradas quantitativamente. Porém, na prática, nem sempre todas as variáveis de um conjunto de dados atendem a esta exigência. Uma solução está em utilizar a ACP com escalonamento ótimo, que é uma técnica que reduz a dimensionalidade, permitindo analisar variáveis observadas em diferentes níveis de mensuração (capítulo 4).

Em um grande conjunto de dados, com muitas observações (n grande) e variáveis (m grande), com diferentes níveis de mensuração (ordinal, nominal, binária e intervalar), algumas tarefas de mineração de dados, como segmentação ou agrupamento, associação e sumarização, podem requerer alto tempo computacional. A técnica da ACP com escalonamento ótimo, é uma alternativa para pré-processar esta grande base de dados, obtendo-se um novo conjunto de k coordenadas. Esta proposta de pré-processamento, considerando variáveis em diferentes níveis de mensuração, está resumida na Figura 5.7.

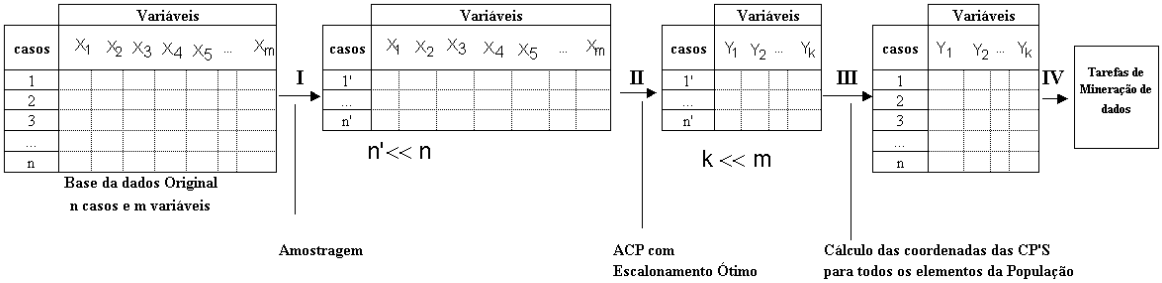

Figura 5.7 Pré-processamento utilizando a ACP com Escalonamento Ótimo

A seguir serão descritas as etapas do pré-processamento enfatizando as fases descritas na Figura 5.7.

Etapa I - Amostragem

Em grandes bases de dados, onde as variáveis podem ser mensuradas em diferentes níveis, a execução do algoritmo de escalonamento ótimo pode requerer muito tempo computacional, uma vez que é um algoritmo iterativo. Uma alternativa, é utilizar amostragem da base de dados, para obter os valores ótimos das categorias. Caso não haja problema com o tempo computacional na execução do escalonamento ótimo, esta etapa pode ser desprezada.

Existem, basicamente, duas formas de amostragem: amostra não-probabilística e probabilística. A forma mais utilizada na mineração de dados é a amostragem probabilística. A amostra probabilística é aquela no qual os elementos da amostra são escolhidos com base em probabilidades conhecidas. Os tipos de amostragem

probabilísticas mais comuns e empregados em mineração de dados são (FERNANDEZ, 2003, p.18):

- Amostragem aleatória simples (AAS)
- Amostragem aleatória estratificada (AAE)
- Amostragem por conglomerados

A Amostragem aleatória simples (AAS) é o método mais comum de amostragem em mineração de dados. Para que este método seja empregado é necessário ter uma listagem de todos os elementos da população (n). Utilizando-se um procedimento aleatório (Tabela de números aleatórios, *software*, ...) sorteia-se, com igual probabilidade, um elemento da população. A seleção da amostra de tamanho n´ pode ser realizada a partir de um *software* estatístico: SPSS, Statistica, SAS entre outros.

A AAS pode ser realizada com reposição, se for permitido que um elemento possa ser sorteado mais de uma vez, e sem reposição, se o elemento for removido da população após ser selecionado.

A Amostragem aleatória estratificada (AAE) consiste em dividir a base de dados em subgrupos ou sub-populações, denominadas estratos. Amostras aleatórias são obtidas de cada estrato. A estratificação é usada principalmente para resolver problemas como:

- a melhoria da precisão das estimativas;
- produzir estimativas para toda a população e subpopulações;
- por questões administrativas entre outras.

Os estratos devem ser internamente mais homogêneos do que a população toda, em relação às variáveis em estudo. A obtenção de estratos mais homogêneos está ligada a um critério de estratificação, sendo assim fundamental um conhecimento prévio sobre a população.

A Amostragem por conglomerados é um agrupamento de elementos da população. Neste método de amostragem, num primeiro estágio a base de dados é dividida em grupos ou conglomerados, e, aleatoriamente, são selecionados alguns desses grupos. Em geral, as bases de dados já se encontram naturalmente divididas em conglomerados.

Num segundo estágio, ou todas as observações dos grupos aleatoriamente selecionados são incluídos no estudo (amostragem de conglomerados em um estágio) ou faz-se uma nova seleção, tomando amostras de elementos dos conglomerados extraídos no primeiro estágio (amostragem de conglomerados em dois estágios).

O tipo de amostragem a ser empregado vai depender do objetivo da mineração de dados. A amostragem aleatória simples é a maneira mais fácil de seleção de uma amostra probabilística de uma população, quando essa corresponde a um arquivo de dados digital. A amostragem por conglomerados tende a produzir resultados menos precisos, maior variância e maiores problemas para análises estatísticas, quando comparada com uma amostra aleatória simples de mesmo tamanho. Sua vantagem pode estar no custo financeiro mais baixo (BARBETTA, 2001,p.52 e BOLFARINE e BUSSAB, 2005, p.160).

Porém, se os dados que serão analisados já se encontram em uma base digital, isto não é relevante. A amostragem estratificada pode produzir resultados mais precisos, mas é necessário algum conhecimento da base para realizar a estratificação de maneira eficiente.

Outro aspecto a ser levado em consideração, é em relação ao tamanho da amostra. O número de variáveis, modelo de análise dos dados (linear, não linear, modelos com interações, e outros) e tamanho da base de dados podem influenciar no tamanho da amostra. Um *default* do SAS *Enterprise Miner* é utilizar 2000 observações obtidas através de uma amostra aleatória simples (FERNANDEZ, 2003, p.18).

Etapa II - ACP com escalonamento ótimo

Esta etapa consiste em executar a ACP com escalonamento ótimo na amostra selecionada anteriormente. O objetivo da técnica é determinar os valores ótimos para as categorias das variáveis e reduzir a dimensionalidade. O algoritmo da ACP com escalonamento ótimo foi descrito na seção 4.2.

Os resultados obtidos na amostra (autovalores, variância explicada, autovetores e número de componentes) são analisados e extrapolados para a população. Os autovalores fornecem o número k de componentes, sendo $k < m$, que são retidos na análise. Para a seleção do número k de autovalores podem ser utilizados os critérios

apresentados na seção 3.7. Os autovetores, que fornecem os coeficientes das componentes principais (seção 3.8), são utilizados para calcular as novas coordenadas para os elementos da população (expressão 3.11).

A execução do algoritmo da ACP com escalonamento ótimo pode ser feita através do *Software* SPSS utilizando a amostra aleatória obtida na etapa anterior. Para tarefas de classificação, em que os dados estão divididos em subgrupos, necessita-se de ajuste para o cálculo das componentes, conforme descrito na seção 3.5.

Etapa III - Cálculo das coordenadas das CP´s para a população

Esta etapa consiste em calcular as novas coordenadas, segundo as k componentes principais obtidas na etapa anterior. Com o objetivo de reduzir a dimensionalidade, o ideal é ter k < m.

Seja Z = $\left(z_{ij}\right)_{nxm}$ a matriz dos dados da base, padronizados, com dimensão n x m representada por

$$Z = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} & \ldots & z_{1m} \\ z_{21} & z_{22} & z_{23} & z_{24} & \ldots & z_{2m} \\ z_{31} & z_{32} & z_{33} & z_{34} & \ldots & z_{3m} \\ \ldots & \ldots. & \ldots.. & \ldots. & \ldots.. & \\ z_{n1} & z_{n2} & z_{n3} & z_{n4} & \ldots & z_{nm} \end{bmatrix}$$

e a matriz V = $\left(v_{jk}\right)_{mxk}$ dos autovetores, de dimensão m x k, representada por

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & v_{13} & \ldots & v_{1k} \\ v_{21} & v_{22} & v_{23} & \ldots & v_{2k} \\ v_{31} & v_{32} & v_{33} & \ldots & v_{3k} \\ \ldots & \ldots. & \ldots.. & \ldots.. & \\ v_{m1} & v_{m2} & v_{m3} & \ldots & v_{mk} \end{bmatrix}$$

As novas coordenadas podem ser calculadas através do produto matrical de Z por V, resultando numa matriz Y = $(y_{ik})_{nxk}$, com dimensão n x k, conforme já discutido na seção 3.8. A representação matricial de Y será:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \ldots. y_{1k} \\ y_{21} & y_{22} & \ldots. y_{2k} \\ \ldots. & \ldots.. & \ldots. \\ y_{n1} & y_{n2} & \ldots. y_{nk} \end{bmatrix},$$

sendo que os elementos y_{ik} da matriz e Y = $(y_{ik})_{nxk}$, resultantes da multiplicação das matrizes Z e V , podem ser calculados por:

$$y_{11} = z_{11}v_{11} + z_{12}v_{21} + z_{13}v_{31} + \ldots + z_{1m}v_{m1}$$
$$y_{12} = z_{11}v_{12} + z_{12}v_{22} + z_{13}v_{32} + \ldots + z_{1m}v_{m2}$$
$$\ldots.$$
$$y_{1k} = z_{11}v_{1k} + z_{12}v_{2k} + z_{13}v_{3k} + \ldots + z_{1m}v_{mk}$$
$$\ldots.$$
$$y_{21} = z_{21}v_{11} + z_{22}v_{21} + z_{23}v_{31} + \ldots + z_{2m}v_{m1}$$
$$y_{2k} = z_{21}v_{1k} + z_{22}v_{2k} + z_{23}v_{3k} + \ldots + z_{2m}v_{mk}$$
$$\ldots.$$
$$y_{n1} = z_{n1}v_{11} + z_{n2}v_{21} + z_{n3}v_{31} + \ldots + z_{nm}v_{m1}$$
$$\ldots$$
$$\ldots$$
$$y_{nk} = z_{n1}v_{1k} + z_{n2}v_{2k} + z_{n3}v_{3k} + \ldots + z_{nm}v_{mk}$$

Aplicando a regra da linha-por-coluna para calcular ZV (LAY, 1999, p.96), tem-se que o produto de Z = $(z_{ij})_{nxm}$ por V = $(v_{jk})_{mxk}$ é a matriz Y = $(y_{ik})_{nxk}$, tal que o elemento y_{ik} é a soma dos produtos da i-ésima linha de Z pelos elementos correspondentes da j-ésima coluna de V.

$$\mathbf{Y} = \mathbf{ZV}$$

$$y_{ij} = \sum_{j=1}^{n} z_{ij}.v_{jk} = z_{i1}v_{1k} + z_{i2}v_{2k} + z_{i3}v_{3k} + \ldots + z_{in}v_{nk} \qquad (5.1)$$

Etapa IV - Tarefas de mineração de dados

Esta última etapa consiste na aplicação de tarefas de mineração de dados (apresentadas na seção 2.2 e exemplificada na seção 3.13) sobre o novo conjunto de dados de menor dimensionalidade, identificado pela matriz Y. Essas tarefas podem estar relacionadas ao alto tempo computacional, que através do novo conjunto de coordenadas poderiam ser executadas em menor tempo.

Uma tarefa muito comum na mineração de dados e que será utilizada para ilustrar a aplicação é a tarefa de agrupamento.

Conforme CHEN *et al.* (1996), Análise de Agrupamentos (AA) é um processo de classificação não supervisionada para agrupar de forma física ou abstrata objetos similares em classes (também chamados de grupos, agrupamentos, conglomerados ou *clusters*).

Segundo HAN e KAMBER (2001), a Análise de Agrupamentos vem sendo largamente utilizada na mineração de dados, decorrente de seus benefícios:

- possibilita ao usuário encontrar grupos úteis;
- auxilia no entendimento das características do conjunto de dados;
- pode ser usada na geração de hipóteses;
- permite predição com base nos grupos formados;
- permite o desenvolvimento de um esquema de classificação para novos dados.

A divisão de um grupo em outros menores se dá através de algoritmos de Análise de Agrupamento. Esses algoritmos basicamente realizam duas tarefas: medem a similaridade entre os objetos e realizam a divisão dos grupos, seguindo um método de formação do agrupamento.

A medida da similaridade está relacionada a um critério que meça a parecença entre dois objetos, ou seja, um critério que diga o quanto dois objetos são parecidos ou não. Para cada tipo de variável existe uma ou mais medidas de similaridade a ser aplicada. Por exemplo, em variáveis intervalares, uma medida de distância muito utilizada é a distância Euclidiana. Para variáveis nominais e ordinais ajustes e transformações das variáveis são necessários e que são discutidos por BUSSAB *et al*. (1990, p.23-39).

Quanto ao método de formação do agrupamento, os algoritmos podem ser classificados em:

- Métodos hierárquicos
- Métodos de partição
- Métodos baseados em modelos
- Métodos baseados em grades
- Métodos baseados em densidade.

Entre os métodos citados, o mais conhecido é o método de partição, utilizando o algoritmo k-médias. Este algoritmo foi proposto por J. MacQueen em 1967. Neste trabalho será utilizado o algoritmo k-médias para a separação dos grupos, por ser um algoritmo muito utilizado e que apresentou um dos melhores resultados no estudo comparativo realizado por PRASS (2004). Uma explicação detalhada quanto ao funcionamento dos métodos de formação de agrupamentos podem ser lida em PRASS (2004, p.32-46).

Outro aspecto importante a ser decidido numa análise de agrupamentos é em relação ao número de grupos em que será dividida a população. No método de partição utilizando o algoritmo k-médias, um critério baseia-se na análise de variância para determinar o número de grupos (BUSSAB *et. al.*, 1990, p.100-102). Outro critério para determinar o número de grupos é a estatística Lambda de Wilks (λ). A estatística λ varia no intervalo [0,1], sendo que quanto mais próximo de zero indica uma melhor separação dos grupos (MANLY, 2005, 46-49). No capítulo 6 será apresentado um exemplo ilustrando a Análise de Agrupamentos.

5.4 Situações de aplicação da proposta

As situações em que a proposta metodológica apresentada é mais indicada:

- pré-processamento em mineração de dados para reduzir a dimensionalidade dos dados, podendo diminuir o tempo computacional para a aplicação de outras técnicas (análise de agrupamentos, classificação, redes neurais, entre outras);

- pré-processamento para aplicação de técnicas de mineração de dados não apropriadas para variáveis não quantitativas;

Exemplo: Um algoritmo muito utilizado na análise de agrupamentos, o k-médias, exige que as variáveis de análise sejam numéricas ou binárias (BERRY e LINOFF, 2004, p.359). Uma solução para este problema é utilizar o escalonamento ótimo que permite gerar novas variáveis quantitativas, a partir de variáveis de diferentes níveis de mensuração (MEULMANN, 2000 p.2.).

- gerar variáveis não correlacionadas. Exemplo: O algoritmo de retropropagação (*backpropagation)* é um algoritmo de aprendizagem muito utilizado em Redes Neurais. Este algoritmo tem sua performance melhorada se as variáveis não forem correlacionadas (HAYKIN,1999, p.208). A ACP transforma variáveis correlacionadas, em novas variáveis, não correlacionadas.

6. APLICAÇÃO

A aplicação conta com a exploração de uma grande base de dados, cuja tarefa da mineração de dados é realizar uma análise de agrupamento, caracterizando grupos similares de pessoas. O arquivo de dados a ser analisado contém registros de pessoas e domicílios pesquisados pelo IBGE no censo 2000.

O presente arquivo de dados contempla situações freqüentemente encontradas nas bases de dados, tais como:

- possuir muitas variáveis, observadas em diferentes níveis de mensuração;
- muitas observações;
- valores faltantes (*missing values*), valores discrepantes e incoerências.

O pré-processamento é uma etapa importante da mineração de dados para garantir a qualidade das análises, tratando as situações descritas anteriormente. Outro aspecto importante do pré-processamento pode ser na redução da dimensionalidade. A base de dados apresenta um total de 91 variáveis com 179.280 registros, observadas em 10 cidades catarinenses, conforme apresentadas na Tabela 6.1.

Tabela 6.1 Base do registros de Pessoas

Cidade	Número de Observações
1. Chapecó	15.031
2. Blumenau	26.578
3. Criciúma	17.060
4. Itajaí	15.071
5. Jaraguá	10.875
6. Joinville	42.641
7. Lages	15.106
8. Palhoça	10.599
9. São José	17.402
10. Tubarão	8.917
TOTAL	179.280

Fonte: IBGE - Censo 2000

Algumas variáveis são referentes às pessoas, outras aos domicílios e as demais referem-se a região do respondente. Para limitar o estudo foram selecionadas 13 variáveis (quantitativas e qualitativas) relacionadas às pessoas, com 18 anos ou mais. A limitação pela idade deve-se ao fato de que as variáveis relacionadas ao trabalho (horas trabalhadas na semana, renda, contribuição à previdência) não se aplicariam para pessoas abaixo de 18 anos. A aplicação desta restrição diminuiu a base de dados para 118.776 registros, conforme apresentado na Tabela 6.2.

Tabela 6.2 Base registros de Pessoas - IBGE - Censo 2000

Cidade	Número de observações (pessoas com 18 anos ou mais)
1. Chapecó	9.534
2. Blumenau	18.217
3. Criciúma	11.121
4. Itajaí	9.877
5. Jaraguá	7.445
6. Joinville	28.454
7. Lages	9.628
8. Palhoça	6.662
9. São José	11.615
10. Tubarão	6.223
TOTAL	118.776

Na seqüência será feito um estudo na base de dados, contemplando desde a análise exploratória, analisando as variáveis e observações através de gráficos e tabelas de freqüência.

Seguindo as etapas de pré-processamento sugeridas na Figura 5.7 a base será pré-processada, a fim de que seja realizada alguma tarefa da mineração de dados. Nas novas variáveis da base pré-processada será possível aplicar técnicas quantitativas, uma vez que o escalonamento ótimo realizou o trabalho de gerar variáveis quantitativas. Finalizando, na base pré-processada, buscar-se-á descrever grupos similares de pessoas, uma tarefa comum em mineração de dados. Para descrever os grupos será utilizado o algoritmo k-médias. Este algoritmo exige que as variáveis de entrada sejam quantitativas, problema já resolvido através do pré-processamento da ACP com EO.

As variáveis selecionadas para análise estão no Quadro 6.1

Quadro 6.1 Variáveis selecionadas

VARIÁVEL	TIPO	DESCRIÇÃO	VALORES e CATEGORIAS
1. IdadeAnos	Intervalar	IDADE (EM ANOS COMPLETOS)	
2. CapEnx	Ordinal	CAPACIDADE DE ENXERGAR	1- incapaz 2- grande dificuldade permanente 3- alguma dificuldade permanente 4- nenhuma dificuldade 9- ignorado *(missing)*
3. CapOuv	Ordinal	CAPACIDADE DE OUVIR	1- incapaz 2- grande dificuldade permanente 3- alguma dificuldade permanente 4- nenhuma dificuldade 9- ignorado *(missing)*
4. LerEscrev	Binária	SABE LER E ESCREVER	1-sim 2-não
5. FreqEscola	Nominal	FREQÜENTA ESCOLA	1- sim, rede particular 2- sim, rede pública 3- não, já freqüentou 4- nunca freqüentou
6. AnosEstudo	Intervalar	TEMPO DE ESTUDO (em anos)	00 - Sem instrução ou menos de 1 ano 01 - 1 ano 02 - 2 anos 03- 3 anos 04 - 4 anos 05 - 5 anos 06 - 6 anos 07 - 7 anos 08 - 8 anos 09 - 9 anos 10 - 10 anos 11 - 11 anos 12 - 12 anos 13 - 13 anos 14 - 14 anos 15 - 15 anos 16 - 16 anos 17 - 17 anos ou mais 20 - Não determinado *(missing)* 30 - Alfabetização de adultos *(missing)*
7. EstCivil	Nominal	ESTADO CIVIL	1- casado(a) 2- desquitado(a) ou separado(a) judicialmente 3- divorciado(a) 4- viúvo(a) 5- solteiro(a)
8. QtsTrabSemana	Ordinal	QUANTOS TRABALHOS TINHA NA SEMANA DE 23 A 29 DE JULHO DE 2000	0 – Não 1- um 2- dois ou mais
9. TrabEra	Nominal	ESSE TRABALHO ERA: ...	0 – Não se aplica (não trabalhou) 1- trabalhador doméstico com carteira de trabalho assinada 2- trabalhador doméstico sem carteira de trabalho assinada 3- empregado com carteira de trabalho assinada 4- empregado sem carteira de trabalho assinada 5- empregador 6- conta-própria 7- aprendiz ou estagiário sem remuneração 8- não remunerado em ajuda a membro do domicílio 9- trabalhador na produção para o próprio consumo
10. ContribInstPrevOf	Nominal	ERA CONTRIBUINTE DE INSTITUTO DE PREVIDÊNCIA OFICIAL	0 – Não se aplica 1-sim 2-não
11. HorasTrab	Ordinal	TOTAL DE HORAS TRABALHADAS	0 – Não trabalhou 1 – até 30 horas 2 – acima de 30 e até 44 horas 3 – acima de 44 horas
12. Aposent	Binária	EM JULHO DE 2000, ERA APOSENTADO DE INSTITUTO DE PREVIDÊNCIA OFICIAL	1-sim 2-não
13. TotRend	Intervalar	TOTAL DE RENDIMENTOS (R$)	

Com a finalidade de agilizar a execução dos algoritmos de agrupamentos, ainda pode ser útil uma redução do número de variáveis. Esta redução será obtida através da análise componentes principais (ACP) com escalonamento ótimo (EO). Na base pré-processada, poderão ser utilizadas técnicas aplicadas às variáveis quantitativas, o que não seria possível com variáveis originais.

6.1 Análise exploratória

Conforme discutido na seção 5.2, para um melhor conhecimento da base de dados, é recomendável iniciar o estudo através da análise exploratória dos dados. Todas as variáveis foram analisadas através de distribuição de freqüências e gráficos apropriados. Os resultados da análise exploratória para cada variável do Quadro 6.1 são apresentados no Anexo 2.

A análise exploratória foi importante para identificar:

- a falta de valores para as variáveis EstCivil e Aposent (3051 observações)
- observações com código 9 (missing value) para as variáveis CapEnx e CapOuv (882 observações)
- observações com idade (IdadeAnos) acima de 100 anos.
- observações com tempo de estudo (AnosEstudo) não determinado e alfabetização de adultos.
- observações com valores bastante altos para renda (TotRenda), tais como de R$ 500.350,00 , R$ 430.000,00 , R$ 200.000,00.

6.2 Limpeza da base de dados

Para os casos com registros faltantes, *missing values*, foi optado por excluir da análise esses registros. Os casos com idade acima de 100 anos e renda acima de R$ 50.000,00 foram considerados como discrepantes e também foram excluídos da análise.

Após esta limpeza, a base passou a ter 113.900 observações, conforme apresentada na Tabela 6.3.

Tabela 6.3 Base registros de Pessoas - IBGE - Censo 2000 (após limpeza)

Cidade	Número de observações
1. Chapecó	9.215
2. Blumenau	17.912
3. Criciúma	10.980
4. Itajaí	9.679
5. Jaraguá	7.289
6. Joinville	28.050
7. Lages	9.525
8. Palhoça	6.607
9. São José	9.495
10. Tubarão	5.148
TOTAL	113.900

A limpeza dos dados (*data cleaning*) foi discutida na seção 2.4, e é um aspecto geral do pré-processamento, e que deve ser considerado na mineração de dados.

6.3 Transformação dos dados

Outra técnica relacionada a aspectos gerais do pré-processamento é a transformação de dados (*data transformation*), e que foi empregada nesta base. A variável renda foi substituída pelo seu logaritmo natural (ln). Devido aos casos com renda é nula, foi somado 1 (um) para calcular o ln, uma vez que ln 0 não é definido. Essa transformação logarítmica foi utilizada para minimizar a assimetria da distribuição, conforme discutido nas seções 2.4.3 e 5.2. O histograma da variável transformada é apresentado na Figura 6.1

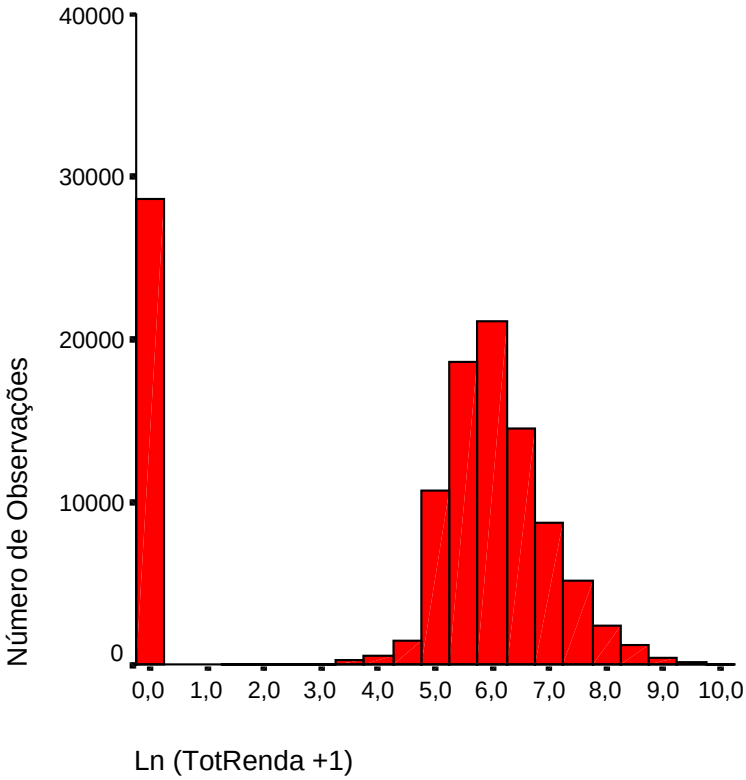

Figura 6.1 Histograma da variável LnRenda

Analisando o histograma da Figura 6.1, observa-se a primeira coluna da esquerda, com mais de 30.000 observações. Essas observações referem-se as pessoas com renda nula.

6.4 Amostragem

Para executar o escalonamento ótimo foi necessário utilizar a amostragem, pois devido à base de dados conter muitas observações, o algoritmo não apresentava resposta. A amostra utilizada foi de 2.225 observações, estratificada proporcionalmente por município. Apesar de não ter sido encontrado nada muito específico na literatura sobre amostragem em mineração de dados, considerou-se uma amostra de 2.225 registros como sendo suficiente. O número de observações selecionadas por cidade está apresentado na Tabela 6.4.

Tabela 6.4 Número de observações amostrada por Cidade

Cidade	Número de Observações	Tamanho da amostra
1. Chapecó	15.031	202
2. Blumenau	26.578	365
3. Criciúma	17.060	193
4. Itajaí	15.071	204
5. Jaraguá	10.875	132
6. Joinville	42.641	543
7. Lages	15.106	172
8. Palhoça	10.599	131
9. São José	17.402	192
10. Tubarão	8.917	91
TOTAL	179.280	2.225

6.5 ACP com escalonamento ótimo

O algoritmo realiza duas tarefas simultâneas: obtém os valores ótimos e reduz a dimensionalidade. Como a base de dados contem mais de 100.000 registros, foi analisada a amostra obtida anteriormente, e na etapa seguinte os resultados foram extrapolados para a população.

A ACP com Escalonamento Ótimo foi realizada no SPSS com as 13 variáveis do Quadro 6.1, e o nível de mensuração utilizado para cada variável está na Tabela 6.5.

Tabela 6.5 Nível de Mensuração do Escalonamento Ótimo

Variável	Tipo	Nível do Escalonamento Ótimo
1. IdadeAnos	Intervalar	Numérico
2. CapEnx	Ordinal	Ordinal
3. CapOuv	Ordinal	Ordinal
4. LerEscrev	Binária	Nominal
5. FreqEscola	Nominal	Nominal
6. AnosEstudo	Intervalar	Numérico
7. EstCivil	Nominal	Nominal
8. QtsTrabSemana	Ordinal	Ordinal
9. TrabEra	Nominal	Nominal
10. ContribInstPrevOf	Nominal	Nominal
11. HorasTr	Ordinal	Ordinal
12. Aposent	Binária	Nominal
13. TotRenda	Intervalar	Numérico

Os valores obtidos do escalonamento ótimo para cada variável qualitativa estão apresentados no Anexo 3.

Os autovalores obtidos através do SPSS da ACP com Escalonamento Ótimo na amostra são apresentados na Tabela 6.6.

Tabela 6.6 Autovalores da Amostra

λ_i	Autovalor	Variância Explicada	Variância Explicada (acumulada)
1	3,76	28,91	28,91
2	2,15	16,51	45,42
3	1,20	9,22	54,64
4	1,12	8,58	63,22
5	0,94	7,25	70,47
6	0,84	6,43	**76,90**
7	0,79	6,07	82,97
8	0,74	5,67	88,64
9	0,61	4,73	93,37
10	0,44	3,35	96,72
11	0,20	1,55	98,27
12	0,12	0,90	99,17
13	0,11	0,83	100,00

Adotando o critério da variância explicada (seção 3.7.3) para a escolha do número de componentes, segundo a Tabela 6.6, pode-se observar que os 6 maiores autovalores explicam 76,90% da variabilidade original. Os autovetores associados aos 6 maiores autovalores estão apresentados na Tabela 6.7 e são os coeficientes das componentes principais.

Tabela 6.7 Autovetores associados aos 6 maiores autovalores

Variável	Dimensão					
	1	2	3	4	5	6
IdadeAnos	-0,1388	-0,5160	-0,1261	0,0904	0,0936	0,0723
CAPENX	0,0755	0,2715	0,2387	0,4599	0,3860	0,0375
CAPOUV	0,0656	0,2499	0,3398	0,4203	0,2670	0,2513
LerEscrev	-0,0605	-0,1327	0,4406	-0,4221	0,5398	-0,4645
FreqEscola	-0,0151	-0,3278	0,3763	0,3250	-0,2535	-0,0067
AnosEstudo	0,1858	0,2924	-0,4525	0,2129	0,0010	-0,2360
ESTCIVIL	0,0131	0,3634	-0,1323	-0,4321	0,2783	0,4260
QtsTrabSemana	0,4808	-0,0399	-0,0160	-0,0243	0,0229	-0,1466
TRABERA	0,4725	-0,1141	0,0911	-0,0677	-0,0554	0,1126
ContribInstPrevOf	0,2897	-0,1514	0,2498	-0,2212	-0,1257	0,5661
HorasTr	0,4729	-0,0631	0,0122	-0,0371	-0,0017	-0,1451
APOSENT	0,1412	0,3992	0,3386	-0,1446	-0,4844	-0,3063
LnRenda	0,3876	-0,2326	-0,2606	0,1070	0,2893	-0,0881

6.6 Cálculo das coordenadas das CP´s para a população

A primeira componente principal com a direção de máxima variabilidade fornece a primeira coordenada Y_1. Sua equação é dada por:

$Y_1 =$ (Z_IdadeAnos **x** –0,1388) + (Z_CAPENX **x** 0,0755) + (Z_CAPOUV **x** 0,0656) + (Z_LerEscrev **x** -0,0605) + (Z_FreqEscola **x** –0,0151) + (Z_AnosEstudo **x** 0,1858) + (Z_ESTCIVIL **x** 0,0131) + (Z_QtsTrabSemana **x** 0,4808) + (Z_TRABERA **x** 0,4725) + (Z_ContribInstPrevOf **x** 0,2897) + (Z_HorasTr **x** 0,4729) + (Z_APOSENT **x** 0,1412) + (Z_LnRenda **x** 0,3876) (6.1)

Foram utilizados os valores padronizados (Z) para as variáveis. Para a primeira observação, aplicando-se (6.1), tem-se:

Y_1 = (-1,35541 **x** –0,1388) + (0,34099 **x** 0,0755) + (0,20351 **x** 0,0656) + (-0,21662 **x** -0,0605) + (-3,25343**x** –0,0151) + (0,93109**x** 0,1858) + (1,22348 **x** 0,0131) + (-1,21325 **x** 0,4808) + (-1,08510 **x** 0,4725) + (-0,53966 **x** 0,2897) + (-1,15733 **x** 0,4729) + (0,38412 **x** 0,1412) + (0,98554 **x** 0,3876)

Y_1 = -0,885

As demais coordenadas, $Y_2, Y_3, Y_4, Y_5 \; e \; Y_6$ são obtidas da mesma maneira, considerando como coeficientes os autovetores da Tabela 6.4.

O novo conjunto de coordenadas, $Y_2, Y_3, Y_4, Y_5 \; e \; Y_6$, dado pela matriz **Y**, conforme visto nas expressões (3.10) e (5.1), pode ser utilizado em análises para outras tarefas de mineração de dados, como por exemplo numa análise de agrupamentos que será apresentada na seção seguinte.

6.7 Tarefas de Mineração de Dados

Conforme já discutido na seção 5.3, identificar grupos numa grande base é uma tarefa freqüente em mineração de dados. Através do escalonamento ótimo com ACP as treze variáveis originais foram reduzidas a seis variáveis quantitativas: $Y_2, Y_3, Y_4, Y_5 \; e \; Y_6$, o que possibilita empregar o método de partição utilizando o algoritmo k-médias para a separação dos grupos.

Para identificar o número de grupos necessários para dividir a base, através do *software* STATISTICA foi calculada a estatística Lambda de Wilks (λ), a partir de uma amostra de 2.225 observações, onde foram consideradas de 2 a 10 divisões e os resultados são apresentados na Figura 6.2.

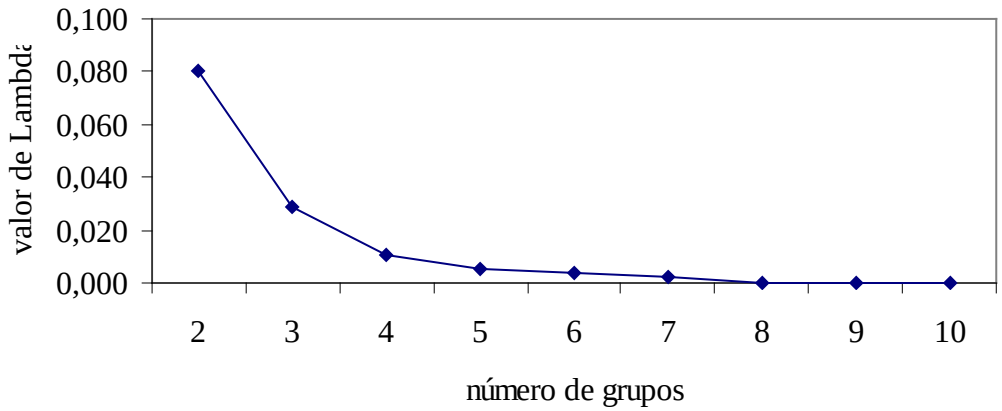

Figura 6.2 Lambda de Wilks para determinar o número de grupos

Analisando a Figura 6.2 pode-se observar que de 2 para 3 e de 3 para 4 grupos tem-se grande redução no valor de Lambda de Wilks. Então no máximo 4 grupos serão necessários para fazer a separação. Também é possível observar que até 6 grupos o valor da estatística diminui, e para 7 grupos o valor se aproxima de zero. Assim, foi optado por dividir a população em 6 subgrupos ou *clusters*.

Na nova base de dados contendo as 113.900 observações e 6 componentes principais, foi aplicada a análise de agrupamentos utilizando o algoritmo k-médias. Para cada observação foi armazenado o *cluster* ao qual pertence. Esta nova variável (coluna) foi transferida para o arquivo com os dados originais, e utilizando a análise exploratória, cada grupo ou *cluster* foi descrito.

As variáveis quantitativas foram analisadas segundo uma Tabela de médias, obtidas para cada *cluster*. Os resultados com relação às médias de cada *cluster* e o seu tamanho estão na Tabela 6.8.

Tabela 6.8 Médias para as variáveis quantitativas

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
tamanho do Cluster	30.499	6.355	4.987	22.200	40.017	9.841
Idade (anos)	36,00	44,53	51,51	26,84	38,04	62,67
Anos Estudo (anos)	6,73	6,94	0,36	5,99	6,27	4,86
Renda (R$)	72,64	822,77	176,46	601,00	905,87	492,67

Para as variáveis qualitativas foram construídas Tabela de freqüências relativas para cada variável, com as categorias da variável observadas em cada *cluster*. Os resultados estão nas Tabela 6.9 a 6.18

Tabela 6.9 Capacidade de Enxergar	**Percentual**					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1 - Incapaz	0,00	0,22	0,42	0,00	0,00	0,51
2 - grande dificuldade	0,41	10,18	5,69	0,01	0,00	10,42
3 - alguma dificuldade	7,73	70,29	19,29	2,72	0,01	26,79
4 - Nenhuma	91,85	19,31	74,59	97,27	99,99	62,29
TOTAL	100,00	100,00	100,00	100,00	100,01	100,00

Tabela 6.10 Capacidade de Ouvir	**Percentual**					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1 - Incapaz	0,00	0,54	0,64	0,00	0,00	0,51
2 - grande dificuldade	0,00	3,29	2,25	0,00	0,00	4,26
3 - alguma dificuldade	1,00	26,53	9,20	0,22	0,00	18,51
4 - Nenhuma	99,00	69,65	87,91	99,78	100,00	76,72
TOTAL	100,00	100,00	100,00	100,01	100,00	100,00

Tabela 6.11 Ler e Escrever	**Percentual**					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1 - Sim	100,00	99,84	0,00	100,00	100,00	98,90
2 - Não	0,00	0,16	100,00	0,00	0,00	1,10
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.12 Freqüenta Escola	**Percentual**					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1 - Sim, Rede Particular	6,08	1,23	0,14	22,03	0,03	0,63
2 - Sim, Rede Pública	11,92	3,12	1,12	14,26	1,23	1,01
3 - Não, já freqüentou	81,44	94,60	30,70	63,71	98,24	95,48
4 - Nunca freqüentou	0,57	1,05	68,04	0,00	0,50	2,89
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.13 Estado Civil

	Percentual					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1 - Casado	49,96	64,83	43,99	9,95	74,72	60,52
2 - Desquitado	2,75	4,88	3,31	1,10	5,02	3,31
3 - Divorciado	1,98	4,37	1,76	1,50	3,57	2,37
4 - Viúvo (a)	4,57	3,41	21,09	0,69	1,96	25,82
5 - Solteiro (a)	40,75	22,50	29,84	86,76	14,73	7,98
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.14 QtsTrabSemana

	Percentual					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0 - Não trabalhou	99,98	0,08	72,51	1,34	0,00	98,73
1 - um	0,02	96,79	26,95	95,64	96,74	1,27
2 - dois	0,00	3,13	0,54	3,02	3,26	0,00
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.15 TrabEra

	Percentual					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0 - Não trabalhou	99,98	0,08	72,51	1,34	0,00	98,73
1 - trab. Doméstico s/	0,02	2,64	1,46	2,82	2,02	0,13
2 - trab. Doméstico c/	0,00	4,85	3,31	1,54	3,91	0,05
3 - empregado c/	0,00	40,44	9,44	71,57	42,71	1,00
4 - empregado s/	0,00	17,32	4,39	14,58	16,33	0,07
5 - empregador	0,00	4,70	0,12	1,74	5,83	0,00
6 - conta própria	0,00	27,55	7,76	4,85	27,84	0,00
7 - estagiário s/ remun.	0,00	0,30	0,00	0,77	0,12	0,00
8 - não remunerado	0,00	1,07	0,48	0,73	0,92	0,00
9 - próprio consumo	0,00	1,04	0,52	0,07	0,32	0,02
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.16 ContribInstPrevOf

	Percentual					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0 - Não se aplica	100,00	49,82	84,78	83,11	49,93	99,93
1 - Sim	0,00	16,08	1,72	6,12	16,91	0,02
2 - Não	0,00	34,10	13,50	10,77	33,16	0,05
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.17 TotTrab	**Percentual**					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0 - Não trabalhou	99,98	0,08	72,51	1,34	0,00	98,73
1 - até 30 horas	0,02	9,63	3,71	7,80	7,66	0,28
2 - Acima de 30 e até 44h	0,00	49,14	13,11	63,60	51,86	0,84
3 - Acima de 44 h	0,00	41,15	10,67	27,25	40,48	0,14
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Tabela 6.18 Aposent	**Percentual**					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1 - Sim	3,83	10,13	34,47	0,78	6,23	85,82
2 - Não	96,17	89,87	65,53	99,22	93,77	14,18
TOTAL	100,00	100,00	100,00	100,00	100,00	100,00

Algumas características observadas nas Tabelas 6.8 a 6.18 para os 6 grupos (*clusters*) são descritas a seguir :

Cluster 1: Segundo maior cluster com 30.499 observações. É formado por pessoas com média de idade de 36 anos e que não trabalham (99,98%). Por este motivo justifica-se a menor renda dos 6 *clusters*. Também pode-se observar que 18% das pessoas deste *cluster* são estudantes de escolas públicas e particulares.

Cluster 2: Com 6.355 observações é formado por pessoas com média de idade de 44 anos e renda média de R$ 822,77. Em relação a variável QtsTrabSemana, a maioria tem um trabalho (96,79%), e são empregados com carteira assinada (40,44%) ou que trabalham por conta própria (27,55%). Também nota-se que um elevado percentual (41,15%) trabalha acima de 44 horas semanais (TotTrab). Em média as pessoas deste grupo estudaram durante 7 anos.

Cluster 3: Menor grupo, formando por 4.987 observações. Este grupo destaca-se dos demais pelo fato de que o tempo médio de estudo é de 0,36 anos e que 100% não sabem ler e escrever. Também é formado por 34,47% de aposentados (Aposent), com renda média de R$ 176,46 e com média de idade de 51,51 anos.

Cluster 4: Este grupo formado por 22.200 observações é o grupo mais jovem, com média de idade de 26,84 anos. 36,29% ainda freqüentam a escola, e a maioria, 86,76% é solteiro (a) (Estado Civil). A maioria trabalha (QtsTrabSemana) e tem um ou dois empregos (98,66%) e 71,57% são empregados com carteira assinada (TrabEra).

Cluster 5: É o maior cluster com 40.017 observações. É formado por pessoas com média de idade de 38,04 anos e que a maioria é casado (a) (74,72%). Também é o *cluster* formado por pessoas com a maior renda (média de R$905,87). A maior porcentagem de empregadores (5,87%) encontra-se neste *cluster*, talvez um motivo para a maior renda (TrabEra). Apresenta muitas pessoas que são empregados com carteira assinada (42,71%) ou que trabalham por conta própria (27,84%).

Cluster 6: Este grupo formado por 9.841 observações é o grupo com maior média de idade: 62,67 anos. Assim, neste grupo predominam pessoas que são aposentadas (85,82%). A maioria é casado (a) (60,52%) e naturalmente apresenta o mais alto percentual de viúvos (as) (25,82%) dos 6 *clusters*.

Uma observação geral que pode ser feita em relação aos *clusters* 6, 3 e 2 é em relação às variáveis capacidade de enxergar e ouvir. Nestas variáveis ocorreu a presença de observações nas categorias alguma dificuldade e grande dificuldade. Isto mostra que a capacidade de enxergar e ouvir está relacionada à idade, pois estes clusters são os que apresentam as maiores médias para a variável idade, principalmente o *cluster* 6 onde o percentual foi mais elevado.

7. CONSIDERAÇÕES FINAIS

7.1 Conclusão

Os procedimentos usuais do pré-processamento incluem várias técnicas, tais como limpeza, transformação, integração e redução dos dados. Tais procedimentos visam garantir qualidade das análises, uma vez que é comum encontrar registros incompletos, valores discrepantes, assimetria, entre outros problemas.

Neste trabalho foi apresentada uma metodologia para o pré-processamento, que vai além dos procedimentos usuais, abordando a mensuração das variáveis de entrada e o tempo computacional.

A metodologia agregou técnicas estatísticas (análise de componentes principais e escalonamento ótimo), e, aplica-se à mineração de dados, permitindo analisar bases que contenham variáveis mensuradas em diferentes níveis, gerando um novo conjunto de variáveis, quantitativas, com dimensionalidade reduzida. Sob o novo conjunto de coordenadas, obtido através do pré-processamento da ACP com EO, poderão ser aplicadas técnicas quantitativas, o que nem sempre é possível nas variáveis originais. Como exemplo, a técnica de agrupamentos, utilizando o algoritmo k-médias, exige que as variáveis de entrada sejam quantitativas (BERRY e LINOFF, 2004, p.359).

A análise de componentes principais é freqüentemente utilizada como uma etapa intermediária de grandes análises, podendo servir como pré-processamento para outras técnicas, como por exemplo, regressão múltipla e análise de agrupamentos.

Segundo MANLY (2005, p.130), alguns algoritmos de análise de agrupamentos começam fazendo uma análise de componentes principais, reduzindo as variáveis originais num número menor de componentes. Segundo o autor, isto pode reduzir drasticamente o tempo computacional para a análise de agrupamentos.

A ACP, como etapa de pré-processamento para outras tarefas, também é indicada para melhorar o desempenho do algoritmo de retropropagação em Redes Neurais (RN).

O algoritmo de retropropagação ou *backpropagation* é um algoritmo de aprendizagem muito utilizado em RN.

Porém, ao utilizar-se a análise de componentes principais com escalonamento ótimo, alguns aspectos deverão ser levados em consideração:

- perda de informação ao reduzir o conjunto de variáveis a poucas componentes;
- substituição das variáveis originais por novas variáveis, as componentes principais, o que dificulta a interpretação.

A metodologia proposta foi testada em uma base com 118.776 observações e 13 variáveis, mensuradas em diferentes níveis, chegando-se aos seguintes resultados:

- Redução de dimensionalidade: Através da ACP com EO as 13 variáveis foram reduzidas a 6 componentes, preservando mais de 76% da variabilidade original;

- Divisão da base em 6 subgrupos ou *clusters*: Sob as novas coordenadas obtidas da ACP com EO foi aplicada a análise de agrupamentos, utilizando o método das k-médias, onde foi possível identificar de forma razoavelmente clara 6 subgrupos ou *clusters*.

Apesar da proposta não ter sido aplicada para outras tarefas de mineração de dados, os resultados obtidos na aplicação foram bem positivos, no sentido de variabilidade explicada (76%) e identificação de subgrupos.

Para problemas de aprendizagem supervisionada, tais como, classificação, predição ou previsão, a metodologia aqui proposta pode precisar de ajustes, que não foram discutidos neste trabalho.

Em relação ao tempo computacional, não foram realizados estudos detalhados, limitando-se a redução de dimensionalidade (redução do número de variáveis em poucas componentes).

Outra limitação do trabalho foi em relação aos dados faltantes, registros incompletos e valores discrepantes. A solução adotada foi excluir estes casos da análise.

7.2 Sugestões de trabalhos futuros

Para pesquisas futuras é indicado:

- Estudo mais aprofundado para se utilizar a ACP com EO como etapa de pré-processamento em situações que envolvem tarefas de classificação e predição (métodos supervisionados);

- Estudo para tratar as observações com valores discrepantes, não se limitando a exclusão do registro;

- Melhor avaliação da relação custo-benefício entre a perda de informação com ACP/EO e as suas vantagens (redução de dimensionalidade, variáveis resultantes não correlacionadas e escalares);

- Implementação de um sistema computacional de pré-processamento que inclua os procedimentos descritos no capítulo 5;

- Comparar o tempo computacional e os resultados encontrados da ACP/EO com a utilização de transformações de variáveis, quando há presença de variáveis em diferentes níveis de mensuração.

REFERÊNCIAS BLIOGRÁFICAS

BANET,T.A.; MORINEAU, A. **Aprender de los datos:** el análisis de componentes principales. España, Barcelona: UEB, 1999.

BARBETTA, Pedro A. **Estatística aplicada às ciências sociais**. Florianópolis: Editora da UFSC, 2001. 4ª edição. p. 69-121.

BELL, Richard. **Lecture 8: Categorical Data Analysys 1: Optimal Scaling.** University of Melbourne, Austrália. Disponível em: < http://www.psych.unimelb.edu.au/staff/bell.html />. Acesso em: 21 junho 2004.

BERRY, Michael J. A.; LINOFF, Gordon S. **Data mining techniques**. USA: Wiley Publishing Inc, 2004. 2ª edição.

BOLFARINE, Heleno, BUSSAB,Wilton O. **Elementos de Amostragem.** São Paulo: Blücher, 2005. 1ª edição. p.15-19; 61-62; 93-95; 159-160.

BUSSAB, Wilton O.; MIAZAKI, Édina S.; ANDRADE, Dalton F. **Introdução à análise de agrupamentos.** 9º Simpósio Brasileiro de Probabilidade e Estatística. São Paulo: ABE, 1990

BUSSAB,Wilton O.; MORETTIN, Pedro A. **Estatística Básica.** São Paulo: Saraiva, 2003. 5ª edição. p.48, 53, 262.

BUSSINES OBJECTS. **Business Intelligence Products**. Informações disponíveis em: <www. businessobjects.com/products/default.asp>. Acesso em: 10 março 2005.

CHATTERJEE, S.; HADI, A. S.; PRICE B. **Regression Analysis By Example.** New York: John Wiley & Sons, Inc., 2000. 3ª. edição. p. 263-272.

CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S.. **Data Mining: An Overview from a Database Perspective**, IEEE Transactions on Knowledge and Data Engineering, v.8, n.6,December 1996.

DTG. Faculty of Social Behavioral Sciences. **Data Theory Group.** Informações disponíveis em: < http://www.datatheory.nl/pages/staff/>. Acesso em: 08 agosto 2004..

FERNANDEZ, G. **Data Mining using SAS applications**. USA: Chapman & Hall, 2003.

GERARDO, Bobby D; LEE, Jaewan; RA, Inho; *at al*. **Association Rule Discovery in Data Mining by Implementing Principal Component Analysis.** Disponível em: < http://www.springerlink.com/(eo00zs550osf5u55k1deg0yd)/app/home/contribution.asp?referrer=parent&backto=searcharticlesresults,3,36;>. Acesso em 31 março 2005.

GNANADESIKAN, R. **Methods for Statical Data Analysis of Multivariate Observation**. USA: John Wiley & Sons, 1977. p. 5-15.

HAN, Jiawei; KAMBER, Micheline. **Data Mining:** concepts and techniques. San Diego: Academic Press, 2001.

HAYKIN, S. **Redes Neurais:** Princípios e prática. São Paulo: Bookman, 1999, 2ª edição. p. 205-209.

IVANQUI, Ivan Ludgero; BARBETTA, Pedro Alberto. **Uso da Análise de Componentes Principais com Escalonamento Ótimo em Escalas tipo Likert.** Simpósio Brasileiro de Probabilidade e Estatística. Minas Gerais: ABE, 2002

JOHNSON,R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. New Jersey: Prentice-Hall, 2002. 5ª edição.

KARGUPTA, H.; HUANG, W.; SIVAKUMAR, K.; *at al*. **Distributed Clustering Using Collective Principal Component Analysis.** Disponível em: <

http://www.springerlink.com/(eo00zs550osf5u55k1deg0yd)/app/home/contribution.asp?referrer=parent&backto=searcharticlesresults,6,36;>. Acesso em 31 março 2005.

LAY, David C. **Álgebra Linear e suas aplicações**.Rio de Janeiro: LTC Editora, 1999, 2ª edição. p.96.

LEBART, L.; MORINEAU, A.; PIRON, M. **Statistique exploratories muldimensionnalle**. Paris: Dumond,1995.

LEEUW, Jan de. **UCLA Departament of Statistics.** Informações disponíveis em : < http://gifi.stat.ucla.edu/>. Acesso em: 10 fevereiro, 2005

LEVINE, D.M.; BERENSON, M. L; STEPHAN, D. **Estatística: Teoria e Aplicações.** Rio de Janeiro: LTC, 2000. p. 219-226; 39-241.

LOESH, Cláudio. **LHSTAT**. Informações Disponíveis em: <www.furb.br/clubevirtual/>. Acesso em: 15 novembro 2004.

MANLY, Bryan F.J. **Multivariate Statiscal Methods**. Chapman & Hall: London, 1995.

MEULMANN Jacqueline J. **Optimal scaling methods for multivariate categorical data analysis**, White Paper – SPSS, Chicago. 2000. Disponível em: <http://whitepapers.zdnet.co.uk/0,39025945,60006271p-39000537q,00.htm>. Acesso em 25 julho, 2005

MEULMANN Jacqueline J.; HEISER Willem J. **SPSS Categories 11.0 -** Manuais do SPSS. Chigago: SPSS Inc, 2001.

PRASS, Fernando S. **Estudo comparativo entre algoritmos de análise de agrupamentos em *data mining*.** Florianópolis, 2004, 71p. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-graduação em Ciência da Computação, UFSC, 2004.

REIS, E. **Estatística Multivariada Aplicada**. Lisboa: Edições Silabo, 1997.

SAS. **SAS.** Informações disponíveis em: < www.sas.com/>. Acesso em: 10 março 2005.

SCREMIN, M.A.A. **Método para a seleção do número de componentes principais com base na lógica difusa**. Florianópolis, 2003, 124p. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-graduação em Engenharia de Produção, UFSC, 2003.

SPSS. **SPSS.** Informações disponíveis em: < www.spss.com/>. Acesso em: 11 março 2005.

STATISTICA. **STATISTICA Products.** Informações Disponíveis em: <www.statsoft.com/products/products.htm/>. Acesso em: 12 março 2005.

TAKANE, Yoshio. **McGill University.** Informações disponíveis em : < http://www.psych.mcgill.ca/labs/lnsc/html/person-yoshio.html>. Acesso em: 10 fevereiro, 2005

YANG, H.; YANG, T. **Outlier Mining Based on Principal Component Estimation.** Disponível em: < http://www.springerlink.com/(eo00zs550osf5u55k1deg0yd)/app/home/contribution.asp?referrer=parent&backto=searcharticlesresults,5,36;>. Acesso em 31 março 2005.

YOUNG, Forrest W. **Quantitative Analysis of Qualitative Data.** Psychometrika – Vol 46, nº 4, dezembro , 1981.

YOUNG, Forrest W; TAKANE,Y. e DE LEEUW, J. **The Principal Components of Mixed Measurement Level Multivariate Data: Alternating Least Squares Method With Optimal Scaling Features**. Psychometrika – Vol 43, nº 2, junho, 1978.

YOUNG, Forrest W. **University of North Carolina.** Informações disponíveis em : < http://forrest.psych.unc.edu/>. Acesso em: 10 fevereiro, 2005.

ANEXOS

Anexo 1 - Funcionamento do algoritmo PRINCIPALS

1. Variáveis Ordinais

Exemplo: Considere um questionário para avaliar o desempenho docente aplicado a 5 alunos, portanto n = 5 observações e m=3 variáveis, sendo que as respostas possíveis em cada pergunta são:

$a_1 =$ ruim ou péssimo $\qquad a_2 =$ regular $\qquad a_3 =$ bom ou ótimo

então $a_1 < a_2 < a_3$ (escala ordinal).

Os resultados dos dados coletados estão apresentados no Quadro 1.

Quadro 1 - Variáveis Ordinais

	Variáveis		
obs.	X_1	X_2	X_3
1	a_3	a_3	a_3
2	a_2	a_1	a_2
3	a_3	a_3	a_2
4	a_2	a_1	a_1
5	a_3	a_2	a_3

As observações 1, 3 e 5 e as observações 2 e 4 formam grupos de alunos similares entre si, pois somente diferem a resposta em uma variável.

Passo 0: Determinar a matriz das observações X_{nxm}

Tem-se que n é o número de observações (casos ou indivíduos) e m é o número de variáveis.

$$\mathbf{X}_{5x3} = \begin{bmatrix} a_3 & a_3 & a_3 \\ a_2 & a_1 & a_2 \\ a_3 & a_3 & a_2 \\ a_2 & a_1 & a_1 \\ a_3 & a_2 & a_3 \end{bmatrix}$$, n = 5 casos e m = 3 variáveis.

Substituindo os elementos por valores arbitrários que mantêm a ordinalidade, por exemplo $a_1 = 1 \quad a_2 = 3 \quad a_3 = 5$, tem-se que:

$$\mathbf{X}_{5x3} = \begin{bmatrix} 5 & 5 & 5 \\ 3 & 1 & 3 \\ 5 & 5 & 3 \\ 3 & 1 & 1 \\ 5 & 3 & 5 \end{bmatrix}$$

Passo 1: Criar a matriz Z, padronizando X por coluna

$$\mathbf{Z}_{5x3} = \begin{bmatrix} 0{,}730 & 1{,}000 & 0{,}956 \\ -1{,}095 & -1{,}000 & -0{,}239 \\ 0{,}730 & 1{,}000 & -0{,}239 \\ -1{,}095 & -1{,}000 & -1{,}434 \\ 0{,}730 & 0{,}000 & 0{,}956 \end{bmatrix}$$

Normalizando a matriz **Z**, para $\mathbf{Z}^*$ tem-se que :

$$\mathbf{Z}^*_{5x3} = \begin{bmatrix} 0{,}365 & 0{,}500 & 0{,}478 \\ -0{,}548 & -0{,}500 & -0{,}120 \\ 0{,}365 & 0{,}500 & -0{,}120 \\ -0{,}548 & -0{,}500 & -0{,}717 \\ 0{,}365 & 0{,}000 & 0{,}478 \end{bmatrix}$$

Passo 2: Estimativa do Modelo

Calcula-se a matriz de correlações **R** a partir de $\mathbf{Z}^*$, através da expressão :

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^*$$

$$\mathbf{R}_{3X3} = \begin{bmatrix} 1 & 0{,}913 & 0{,}764 \\ 0{,}913 & 1 & 0{,}598 \\ 0{,}764 & 0{,}598 & 1 \end{bmatrix}$$

Autovalores de **R**, calculados pela equação característica, $\left| \mathbf{R} - \lambda \mathbf{I} \right| = 0$ são :

$\lambda_1 = 2{,}523425$

$\lambda_2 = 0{,}420476$

$\lambda_3 = 0{,}056100$

Os dois primeiros autovalores dão uma variância explicada de $\frac{2{,}523425 + 0{,}420476}{2{,}523425 + 0{,}420476 + 0{,}056100} = \frac{2{,}944}{3{,}000} = 98{,}13\%$.

Sendo $\mathbf{D}_k$ a matriz diagonal formada pelos k maiores autovalores, considerando k = 2 componentes, tem-se:

$$\mathbf{D}_2 = \begin{bmatrix} 2{,}523 & 0 \\ 0 & 0{,}420 \end{bmatrix} \quad \text{e sua inversa} \quad \mathbf{D}_2^{-1} = \begin{bmatrix} 0{,}396 & 0 \\ 0 & 2{,}378 \end{bmatrix}$$

Sendo **V** a Matriz dos Autovetores associados:

$$\mathbf{V} = \begin{bmatrix} -0{,}615 & 0{,}167 & 0{,}771 \\ -0{,}579 & 0{,}568 & -0{,}585 \\ -0{,}535 & -0{,}806 & -0{,}253 \end{bmatrix}$$

considerando k = 2 componentes , tem-se:

$$\mathbf{V}_2 = \begin{bmatrix} -0{,}615 & 0{,}167 \\ -0{,}579 & 0{,}568 \\ -0{,}535 & -0{,}806 \end{bmatrix}$$

Passo 3: Determinar a Matriz $\mathbf{Z}^{"} = \mathbf{FL}^t$

A matriz $\mathbf{F}_{nxk}$, contém os *n* escores relativos aos *k* componentes principais; e a matriz $\mathbf{L}_{mx\,k}$, contém as cargas fatoriais das *m* variáveis em termos dos *k* componentes.

Cálculos auxiliares para determinar $\mathbf{Z}^{''}$

$$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j$$

$$\mathbf{L} = \begin{bmatrix} -0{,}977 & 0{,}108 \\ -0{,}919 & 0{,}369 \\ -0{,}851 & -0{,}552 \end{bmatrix}$$

$$\mathbf{F} = \mathbf{Z}^{*}\mathbf{L}\mathbf{D}_{\mathrm{K}}^{-1}$$

$$\mathbf{F} = \begin{bmatrix} -0{,}485 & -0{,}062 \\ 0{,}435 & -0{,}431 \\ -0{,}283 & 0{,}681 \\ 0{,}636 & 0{,}312 \\ -0{,}303 & -0{,}500 \end{bmatrix}$$

$$\mathbf{Z}^{''} = \begin{bmatrix} 0{,}467 & 0{,}423 & 0{,}445 \\ -0{,}471 & -0{,}558 & -0{,}145 \\ 0{,}350 & 0{,}511 & -0{,}115 \\ -0{,}588 & -0{,}470 & -0{,}704 \\ 0{,}242 & 0{,}094 & 0{,}519 \end{bmatrix}$$

Verificando a qualidade do ajuste $\theta^{*} = tr\left[\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)\right]$

$$\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right) = \begin{bmatrix} 0{,}033 & -0{,}025 & -0{,}011 \\ -0{,}025 & 0{,}019 & 0{,}008 \\ -0{,}0011 & 0{,}008 & 0{,}004 \end{bmatrix}$$

$$\theta^{*} = 0{,}056$$

Passo 4: Escalonamento Ótimo

Seja o vetor **B**, contendo os valores de $\mathbf{Z}^*$ ordenados crescentemente.

Fazendo o escalonamento por variável, tem-se:

Para a variável A:

$$\mathbf{Z}^* = \begin{bmatrix} 0{,}365 & 0{,}500 & 0{,}478 \\ -0{,}548 & -0{,}500 & -0{,}120 \\ 0{,}365 & 0{,}500 & -0{,}120 \\ -0{,}548 & -0{,}500 & -0{,}717 \\ 0{,}365 & 0{,}000 & 0{,}478 \end{bmatrix}$$

$$\mathbf{B}^t = \begin{bmatrix} -0{,}548 & -0{,}548 & 0{,}365 & 0{,}365 & 0{,}365 \end{bmatrix}.$$

Seja o vetor $\hat{\mathbf{Z}}^t$ contendo os elementos de $\mathbf{Z}^{''}$

(Os elementos de $\hat{\mathbf{Z}}^t$ mantém uma ordem em correspondência com a matriz **B**)

$$\hat{\mathbf{Z}}^{\mathbf{t}} = \begin{bmatrix} -0{,}471 & -0{,}588 & 0{,}467 & 0{,}350 & 0{,}242 \end{bmatrix}$$

Criação da Matriz Indicadora: **G**

$$\mathbf{G}^t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Calcula-se agora $\mathbf{Z}^G = \mathbf{G}\left(\mathbf{G}^t\mathbf{G}\right)^{-1}\mathbf{G}^t\,\hat{\mathbf{Z}}^t$

$$\mathbf{Z}^G = \begin{bmatrix} -0{,}5295 \\ -0{,}5295 \\ 0{,}3530 \\ 0{,}3530 \\ 0{,}3530 \end{bmatrix}.$$

Recolocam-se os resultados numa nova matriz **X** como foi usado para fazer de **Z** para **B**, e recomeça o processo a partir desta nova matriz **X**.

Recolocando:

$$\mathbf{X} = \begin{bmatrix} 0{,}353 & ..... & ..... \\ -0{,}529 & ..... & ..... \\ 0{,}353 & ..... & ..... \\ -0{,}529 & ..... & ..... \\ 0{,}353 & ..... & ..... \end{bmatrix}$$

O mesmo é feito para as demais variáveis, onde vamos obter que:

$$\mathbf{X} = \begin{bmatrix} 0{,}353 & 0{,}467 & 0{,}482 \\ -0{,}529 & -0{,}514 & -0{,}130 \\ 0{,}353 & 0{,}467 & -0{,}130 \\ -0{,}529 & -0{,}514 & -0{,}704 \\ 0{,}353 & 0{,}094 & 0{,}482 \end{bmatrix}$$

A partir desta nova matriz **X** , repete-se o passo 1 ao 3, até que a convergência do método seja alcançada.

Retornando ao **Passo 1**, tem-se:

A partir da matriz **Z**, normalizando obtém-se $\mathbf{Z}^* = \begin{bmatrix} 0{,}365 & 0{,}473 & 0{,}483 \\ -0{,}548 & -0{,}521 & -0{,}130 \\ 0{,}365 & 0{,}473 & -0{,}130 \\ -0{,}548 & -0{,}521 & -0{,}706 \\ 0{,}365 & 0{,}095 & 0{,}483 \end{bmatrix}$

Passo 2: Calcula-se a matriz de correlações **R** a partir de $\mathbf{Z}^*$, através da expressão:

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^*$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}951 & 0{,}764 \\ 0{,}951 & 1 & 0{,}649 \\ 0{,}764 & 0{,}649 & 1 \end{bmatrix}_{3x3}$$

Autovalores de **R**, calculados pela equação característica, $\left| \mathbf{R} - \lambda \mathbf{I} \right| = 0$ são:

$\lambda_1 =$ 2,583 $\lambda_2 = 0{,}383$ $\lambda_3 = 0{,}034$

Os dois primeiros autovalores dão uma variância explicada de $\frac{2{,}583 + 0{,}383}{2{,}583 + 0{,}383 + 0{,}034} = \frac{2{,}966}{3{,}000} = 98{,}86\%$.

Podemos observar que utilizando dois componentes principais, o percentual da variância explicada passou de 98,13% para 98,86%.

Sendo $\mathbf{D}_k$ a matriz diagonal formada pelos k maiores autovalores, considerando k = 2 componentes, tem-se:

$$\mathbf{D}_2 = \begin{bmatrix} 2{,}583 & 0 \\ 0 & 0{,}383 \end{bmatrix} \text{ e sua inversa } \mathbf{D}_2^{-1} = \begin{bmatrix} 0{,}387 & 0 \\ 0 & 2{,}609 \end{bmatrix}$$

Sendo **V** a Matriz dos Autovetores associados: $\mathbf{V} = \begin{bmatrix} -0{,}610 & 0{,}231 & 0{,}758 \\ -0{,}586 & 0{,}514 & -0{,}627 \\ -0{,}534 & -0{,}826 & -0{,}178 \end{bmatrix}$

considerando k = 2 componentes, tem-se: $\mathbf{V}_2 = \begin{bmatrix} -0{,}610 & 0{,}231 \\ -0{,}586 & 0{,}514 \\ -0{,}534 & -0{,}826 \end{bmatrix}$

Passo 3: Determinar a Matriz $\mathbf{Z}^{"} = \mathbf{FL}^{t}$

Cálculos auxiliares para determinar $\mathbf{Z}^{"}$

$$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j$$

$$\mathbf{L} = \begin{bmatrix} -0{,}980 & 0{,}143 \\ -0{,}941 & 0{,}318 \\ -0{,}859 & -0{,}512 \end{bmatrix}$$

$$\mathbf{F} = \mathbf{Z}^{*}\mathbf{L}\mathbf{D}_k^{-1}$$

$$\mathbf{F} = \begin{bmatrix} -0{,}472 & -0{,}116 \\ 0{,}441 & -0{,}463 \\ -0{,}268 & 0{,}703 \\ 0{,}632 & 0{,}306 \\ -0{,}334 & -0{,}430 \end{bmatrix}$$

$$\mathbf{Z}^{"} = \begin{bmatrix} 0{,}446 & 0{,}407 & 0{,}464 \\ -0{,}498 & -0{,}562 & -0{,}142 \\ 0{,}363 & 0{,}475 & -0{,}130 \\ -0{,}576 & -0{,}498 & -0{,}700 \\ 0{,}266 & 0{,}177 & 0{,}507 \end{bmatrix}$$

Verificando a qualidade do ajuste $\theta^{*} = tr\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right)$

$$\left(\mathbf{Z}^{*} - \mathbf{Z}^{t}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right) = \begin{bmatrix} 0{,}020 & -0{,}016 & -0{,}005 \\ -0{,}016 & 0{,}013 & 0{,}004 \\ -0{,}005 & 0{,}004 & 0{,}001 \end{bmatrix}$$

$\theta^{*} = 0{,}034 \longrightarrow$ **diminuiu**

E o processo continua até que a diferença de θ entre uma iteração e outra seja desprezível.

Passo 4: Escalonamento Ótimo

Seja o vetor **B** , contendo os valores de $\mathbf{Z}^{*}$ ordenados.

Fazendo o escalonamento por variável, tem-se:

Para a variável A:

$$\mathbf{Z}^{*} = \begin{bmatrix} 0{,}365 & 0{,}473 & 0{,}483 \\ -0{,}548 & -0{,}521 & -0{,}130 \\ 0{,}365 & 0{,}473 & -0{,}130 \\ -0{,}548 & -0{,}521 & -0{,}706 \\ 0{,}365 & 0{,}095 & 0{,}483 \end{bmatrix}$$

$$\mathbf{B}^{t} = \begin{bmatrix} -0{,}548 & -0{,}548 & 0{,}365 & 0{,}365 & 0{,}365 \end{bmatrix}$$

Seja o vetor $\hat{\mathbf{Z}}^{t}$ contendo os elementos de $\mathbf{Z}^{"}$

(Os elementos de $\hat{\mathbf{Z}}^{t}$ mantém uma ordem em correspondência com a matriz **B**)

$$\hat{\mathbf{Z}}^{t} = \begin{bmatrix} -0{,}498 & -0{,}576 & 0{,}446 & 0{,}363 & 0{,}266 \end{bmatrix}$$

Criação da Matriz Indicadora: **G**

$$\mathbf{G}^{t} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Calcula-se agora $\mathbf{Z}^{G} = \mathbf{G}\left(\mathbf{G}^{t}\mathbf{G}\right)^{-1}\mathbf{G}^{t}\,\hat{\mathbf{Z}}^{t}$

$$\mathbf{Z}^{G} = \begin{bmatrix} -0{,}5370 \\ -0{,}5370 \\ 0{,}3580 \\ 0{,}3580 \\ 0{,}3580 \end{bmatrix}$$

Recolocam-se os resultados numa nova matriz **X** como foi usado para fazer de **Z** para **B**, e recomeça o processo a partir desta nova matriz **X**.

Recolocando:

$$\mathbf{X} = \begin{bmatrix} 0{,}358 & ..... & ..... \\ -0{,}537 & ..... & ..... \\ 0{,}358 & ..... & ..... \\ -0{,}537 & ..... & ..... \\ 0{,}358 & ..... & ..... \end{bmatrix}$$

O mesmo é feito para as demais variáveis, onde vamos obter que:

$$\mathbf{X} = \begin{bmatrix} 0{,}358 & 0{,}441 & 0{,}486 \\ -0{,}537 & -0{,}530 & -0{,}136 \\ 0{,}358 & 0{,}441 & -0{,}136 \\ -0{,}537 & -0{,}530 & -0{,}700 \\ 0{,}358 & 0{,}177 & 0{,}486 \end{bmatrix}$$

A partir desta nova matriz **X** , repete-se o passo 1 ao 3, até que a convergência do método seja alcançada.

Retornando ao **Passo 1**, tem-se:

A partir da matriz **Z**, normalizando obtém-se $\mathbf{Z}^* = \begin{bmatrix} 0{,}365 & 0{,}445 & 0{,}486 \\ -0{,}548 & -0{,}535 & -0{,}136 \\ 0{,}365 & 0{,}445 & -0{,}136 \\ -0{,}548 & -0{,}535 & -0{,}700 \\ 0{,}365 & 0{,}179 & 0{,}486 \end{bmatrix}$

Passo 2: Calcula-se a matriz de correlações **R** a partir de $\mathbf{Z}^*$, através da expressão:

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^*$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}976 & 0{,}763 \\ 0{,}976 & 1 & 0{,}690 \\ 0{,}763 & 0{,}690 & 1 \end{bmatrix}$$

Autovalores de **R**, calculados pela equação característica, $\left| \mathbf{R} - \lambda \mathbf{I} \right| = 0$ são:

$\lambda_1 = 2{,}626 \qquad \lambda_2 = 0{,}356 \qquad \lambda_3 = 0{,}018$

Os dois primeiros autovalores dão uma variância explicada de $\frac{2{,}626 + 0{,}356}{2{,}626 + 0{,}356 + 0{,}018} = \frac{2{,}982}{3{,}000} = 99{,}40\%$.

Podemos observar que utilizando dois componentes principais, o percentual da variância explicada passou de 98,86% para 99,40%.

Sendo $\mathbf{D}_k$ a matriz diagonal formada pelos k maiores autovalores, considerando k = 2 componentes, tem-se:

$\mathbf{D}_2 = \begin{bmatrix} 2{,}626 & 0 \\ 0 & 0{,}356 \end{bmatrix}$ e sua inversa $\mathbf{D}_2^{-1} = \begin{bmatrix} 0{,}381 & 0 \\ 0 & 2{,}808 \end{bmatrix}$

Sendo **V** a Matriz dos Autovetores associados: $\mathbf{V} = \begin{bmatrix} -0{,}605 & 0{,}283 & 0{,}744 \\ -0{,}590 & 0{,}468 & -0{,}658 \\ -0{,}534 & -0{,}837 & -0{,}116 \end{bmatrix}$

considerando k = 2 componentes, tem-se: $\mathbf{V}_2 = \begin{bmatrix} -0{,}605 & 0{,}283 \\ -0{,}590 & 0{,}4684 \\ -0{,}534 & -0{,}837 \end{bmatrix}$

Passo 3: Determinar a Matriz $\mathbf{Z}^{"} = \mathbf{FL}^t$

Cálculos auxiliares para determinar $\mathbf{Z}^{"}$

$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j$

$\mathbf{L} = \begin{bmatrix} -0{,}981 & 0{,}169 \\ -0{,}956 & 0{,}279 \\ -0{,}866 & -0{,}500 \end{bmatrix}$

$\mathbf{F} = \mathbf{Z}^{*}\mathbf{L}\mathbf{D}_k^{-1}$

$\mathbf{F} = \begin{bmatrix} -0{,}459 & -0{,}159 \\ 0{,}444 & -0{,}489 \\ -0{,}254 & 0{,}713 \\ 0{,}630 & 0{,}304 \\ -0{,}362 & -0{,}368 \end{bmatrix}$

$$\mathbf{Z}^{''} = \begin{bmatrix} 0{,}423 & 0{,}394 & 0{,}477 \\ -0{,}518 & -0{,}561 & -0{,}140 \\ 0{,}369 & 0{,}442 & -0{,}136 \\ -0{,}567 & -0{,}518 & -0{,}697 \\ 0{,}293 & 0{,}243 & 0{,}497 \end{bmatrix}$$

Verificando a qualidade do ajuste $\theta^{*} = tr\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)$

$$\left(\mathbf{Z}^{*} - \mathbf{Z}^{t}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right) = \begin{bmatrix} 0{,}010 & -0{,}009 & -0{,}002 \\ -0{,}009 & 0{,}008 & 0{,}001 \\ -0{,}002 & 0{,}001 & 0{,}000 \end{bmatrix}$$

$\theta^{*} = 0{,}018 \longrightarrow$ **diminuiu**

E o processo continua até que a diferença de θ entre uma iteração e outra seja desprezível.

2. Variáveis Nominais

Exemplo: Um questionário aplicado a 5 alunos, perguntando sobre Sexo, Região e a Cor, apresentou o seguinte resultado, dado pelo Quadro 2.

Quadro 2 - Variáveis Nominais

	Variáveis		
obs.	**Sexo** X_1	**Região** X_2	**Cor** X_3
1	Masculino	Norte	Parda
2	Masculino	Norte	Parda
3	Feminino	Sudeste	Branca
4	Feminino	Nordeste	Branca
5	Feminino	Sul	Preta

Tratando-se de variáveis nominais, serão atribuídos números as categorias, sem se preocupar com a ordinalidade. Também os valores atribuídos são arbitrários.

Serão escolhidos os seguintes números:

Variável:	Sexo	Região	Cor
	1 – Masculino	1 – Norte	1 – Preta
	2 – Feminino	2 – Sul	2 – Branca
		3 – Nordeste	3 – Parda
		4 – Sudeste	

Passo 0: Determinar a matriz das observações $\mathbf{X}_{nxm}$

$$\mathbf{X}_{5x3} = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ 2 & 4 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 1 \end{bmatrix}$$

Passo 1: Criar a matriz Z, padronizando X por coluna

$$\mathbf{Z}_{5x3} = \begin{bmatrix} -1{,}095 & -0{,}920 & 0{,}956 \\ -1{,}095 & -0{,}920 & 0{,}956 \\ 0{,}730 & 1{,}381 & -0{,}239 \\ 0{,}730 & 0{,}614 & -0{,}239 \\ 0{,}730 & -0{,}153 & -1{,}434 \end{bmatrix}$$

Normalizando a matriz **Z**, para $\mathbf{Z}^*$ tem-se que:

$$\mathbf{Z}^{*}_{5x3} = \begin{bmatrix} -0{,}548 & -0{,}460 & 0{,}478 \\ -0{,}548 & -0{,}460 & 0{,}478 \\ 0{,}365 & 0{,}690 & -0{,}120 \\ 0{,}365 & 0{,}307 & -0{,}120 \\ 0{,}365 & -0{,}077 & -0{,}717 \end{bmatrix}$$

Passo 2: Estimativa do Modelo

Calcula-se a matriz de correlações **R** a partir de $\mathbf{Z}^{*}$, através da expressão:

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^{*}$$

$$\mathbf{R}_{3x3} = \begin{bmatrix} 1 & 0{,}840 & -0{,}873 \\ 0{,}840 & 1 & -0{,}504 \\ -0{,}873 & -0{,}504 & 1 \end{bmatrix}$$

Autovalores de **R**, calculados pela equação característica, $\left| \mathbf{R} - \lambda\mathbf{I} \right| = 0$ são:

$$\lambda_1 = 2{,}4895 \qquad \lambda_2 = 0{,}4964 \qquad \lambda_3 = 0{,}0142$$

Os dois primeiros autovalores dão uma variância explicada de $\frac{2{,}4895 + 0{,}4964}{3} = \frac{2{,}9858}{3{,}000} = 99{,}53\%$.

Sendo $\mathbf{D}_k$ a matriz diagonal formada pelos k maiores autovalores, considerando k = 2 componentes, tem-se:

$$\mathbf{D}_2 = \begin{bmatrix} 2{,}4898 & 0 \\ 0 & 0{,}4964 \end{bmatrix} \quad \text{e sua inversa} \quad \mathbf{D}_2^{-1} = \begin{bmatrix} 0{,}402 & 0 \\ 0 & 2{,}015 \end{bmatrix}$$

Sendo **V** a Matriz dos Autovetores associados:

$$\mathbf{V} = \begin{bmatrix} -0{,}631 & 0{,}024 & 0{,}775 \\ -0{,}543 & -0{,}727 & -0{,}419 \\ 0{,}554 & -0{,}686 & 0{,}472 \end{bmatrix}$$

considerando k = 2 componentes , tem-se: $\mathbf{V}_2 = \begin{bmatrix} -0,631 & 0,024 \\ -0,543 & -0,727 \\ 0,554 & -0,686 \end{bmatrix}$

Passo 3: Determinar a Matriz $\mathbf{Z}^{"} = \mathbf{FL}^{t}$

A matriz $\mathbf{F}_{nxk}$, contém os *n* escores relativos aos *k* componentes principais; e a matriz $\mathbf{L}_{mx\,k}$, contém as cargas fatoriais das *m* variáveis em termos dos *k* componentes.

Cálculos auxiliares para determinar $\mathbf{Z}^{"}$

$$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j$$

$$\mathbf{L} = \begin{bmatrix} -0,996 & 0,017 \\ -0,857 & -0,512 \\ 0,874 & -0,483 \end{bmatrix}$$

$$\mathbf{F} = \mathbf{Z}^{*}\mathbf{L}\mathbf{D}_{\mathrm{K}}^{-1}$$

$$\mathbf{F} = \begin{bmatrix} 0,545 & -0,009 \\ 0,545 & -0,009 \\ -0,426 & -0,584 \\ -0,294 & -0,188 \\ -0,371 & 0,790 \end{bmatrix}$$

$$\mathbf{Z}^{"} = \begin{bmatrix} -0,543 & -0,463 & 0,481 \\ -0,543 & -0,463 & 0,481 \\ 0,414 & 0,664 & -0,090 \\ 0,289 & 0,348 & -0,166 \\ 0,383 & -0,086 & -0,706 \end{bmatrix}$$

Verificando a qualidade do ajuste $\theta^{*} = tr\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right)$

$$\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{"}\right) = \begin{bmatrix} 0,009 & -0,005 & 0,005 \\ -0,005 & 0,002 & -0,003 \\ 0,005 & -0,003 & 0,003 \end{bmatrix}$$

$$\theta^{*} = 0,014$$

Passo 4: Escalonamento Ótimo

Seja o vetor **B** , contendo os valores de $\mathbf{Z}^*$ ordenados.

Fazendo o escalonamento por variável, tem-se:

Para a variável A:

$$\mathbf{Z}^* = \begin{bmatrix} -0{,}548 & -0{,}460 & 0{,}478 \\ -0{,}548 & -0{,}460 & 0{,}478 \\ 0{,}365 & 0{,}690 & -0{,}120 \\ 0{,}365 & 0{,}307 & -0{,}120 \\ 0{,}365 & -0{,}077 & -0{,}717 \end{bmatrix}$$

$$\mathbf{B}^t = \begin{bmatrix} -0{,}548 & -0{,}548 & 0{,}365 & 0{,}365 & 0{,}365 \end{bmatrix}$$

Seja o vetor $\hat{\mathbf{Z}}^t$ contendo os elementos de $\mathbf{Z}^{''}$

(Os elementos de $\hat{\mathbf{Z}}^t$ mantém uma ordem em correspondência com a matriz **B**)

$$\mathbf{Z}^t = \begin{bmatrix} -0{,}543 & -0{,}543 & 0{,}414 & 0{,}289 & 0{,}383 \end{bmatrix}$$

Criação da Matriz Indicadora: **G**

$$\mathbf{G}^t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Calcula-se agora $\mathbf{Z}^G = \mathbf{G}\left(\mathbf{G}^t\mathbf{G}\right)^{-1}\mathbf{G}^t\,\hat{\mathbf{Z}}^t$

$$\mathbf{Z}^{G} = \begin{bmatrix} -0{,}543 \\ -0{,}543 \\ 0{,}362 \\ 0{,}362 \\ 0{,}362 \end{bmatrix}$$

Recolocam-se os resultados numa nova matriz **X** como foi usado para fazer de **Z** para **B**, e recomeça o processo a partir desta nova matriz **X**

Recolocando:

$$\mathbf{X} = \begin{bmatrix} -0{,}543 & ..... & ..... \\ -0{,}543 & ..... & ..... \\ 0{,}362 & ..... & ..... \\ 0{,}362 & ..... & ..... \\ 0{,}362 & ..... & ..... \end{bmatrix}$$

O mesmo é feito para as demais variáveis, onde vamos obter que:

$$\mathbf{X} = \begin{bmatrix} -0{,}543 & 0{,}463 & 0{,}481 \\ -0{,}543 & -0{,}463 & 0{,}481 \\ 0{,}362 & 0{,}664 & -0{,}128 \\ 0{,}362 & 0{,}348 & -0{,}128 \\ 0{,}362 & 0{,}086 & 0{,}706 \end{bmatrix}$$

A partir desta nova matriz **X** , repete-se o passo 0 ao 3, até que a convergência do método seja alcançada.

Retornando ao Passo 1, tem-se:

A partir da matriz **Z**, normalizando obtém-se $\mathbf{Z}^* = \begin{bmatrix} -0{,}548 & -0{,}463 & 0{,}482 \\ -0{,}548 & -0{,}463 & 0{,}482 \\ 0{,}365 & 0{,}665 & -0{,}128 \\ 0{,}365 & 0{,}348 & -0{,}128 \\ 0{,}365 & -0{,}087 & -0{,}708 \end{bmatrix}$

Calcula-se a matriz de correlações **R** a partir de $\mathbf{Z}^*$, através da expressão:

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^*$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}846 & -0{,}881 \\ 0{,}846 & 1 & -0{,}516 \\ -0{,}881 & -0{,}516 & 1 \end{bmatrix}$$

Autovalores de **R**, calculados pela equação característica, $|\mathbf{R} - \lambda\mathbf{I}| = 0$ são:

$\lambda_1 = 2{,}506 \qquad \lambda_2 = 0{,}485 \qquad \lambda_3 = 0{,}009$

Os dois primeiros autovalores dão uma variância explicada de $\frac{2{,}506 + 0{,}485}{2{,}506 + 0{,}485 + 0{,}09} = \frac{2{,}991}{3{,}000} = 99{,}62\%$.

Utilizando dois componentes principais, o percentual da variância explicada passou de 99,53% para 99,62%.

Sendo $\mathbf{D}_k$ a matriz diagonal formada pelos k maiores autovalores, considerando k = 2 componentes, tem-se:

$\mathbf{D}_2 = \begin{bmatrix} 2{,}506 & 0 \\ 0 & 0{,}485 \end{bmatrix}$ e sua inversa $\mathbf{D}_2^{\,1} = \begin{bmatrix} 0{,}399 & 0 \\ 0 & 2{,}062 \end{bmatrix}$

$$\mathbf{V} = \begin{bmatrix} -0{,}630 & 0{,}026 & 0{,}776 \\ -0{,}544 & -0{,}729 & -0{,}416 \\ 0{,}555 & -0{,}684 & 0{,}473 \end{bmatrix}$$

Sendo **V** a Matriz dos Autovetores associados:

considerando k = 2 componentes, tem-se: $\mathbf{V}_2 = \begin{bmatrix} -0{,}630 & 0{,}026 \\ -0{,}544 & -0{,}729 \\ 0{,}555 & -0{,}684 \end{bmatrix}$

Passo 2: Determinar a Matriz $\mathbf{Z}^{''} = \mathbf{F}\mathbf{L}^{t}$

Cálculos auxiliares para determinar $\mathbf{Z}^{''}$

$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j$

$\mathbf{L} = \begin{bmatrix} -0{,}997 & 0{,}018 \\ -0{,}861 & -0{,}508 \\ 0{,}878 & -0{,}477 \end{bmatrix}$

$\mathbf{F} = \mathbf{Z}^{*}\mathbf{L}\mathbf{D}_k^{-1}$

$\mathbf{F} = \begin{bmatrix} 0{,}546 & -0{,}010 \\ 0{,}546 & -0{,}010 \\ -0{,}419 & -0{,}556 \\ 0{,}310 & -0{,}225 \\ -0{,}364 & 0{,}800 \end{bmatrix}$

$\mathbf{Z}^{''} = \begin{bmatrix} -0{,}545 & -0{,}465 & 0{,}484 \\ -0{,}545 & -0{,}465 & 0{,}484 \\ 0{,}407 & 0{,}642 & -0{,}103 \\ 0{,}305 & 0{,}381 & -0{,}165 \\ 0{,}377 & -0{,}093 & -0{,}701 \end{bmatrix}$

Verificando a qualidade do ajuste $\theta^{*} = tr\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)$

$\left(\mathbf{Z}^{*} \quad \mathbf{Z}^{t}\right)^{t}\left(\mathbf{Z}^{*} \quad \mathbf{Z}^{''}\right) = \begin{bmatrix} 0{,}006 & -0{,}003 & 0{,}003 \\ -0{,}003 & 0{,}002 & -0{,}002 \\ 0{,}003 & -0{,}002 & 0{,}002 \end{bmatrix}$

$\theta^{*} = 0{,}010 \longrightarrow$ **diminuiu**

E o processo continua até que a diferença de θ entre uma iteração e outra seja desprezível.

O resultado gráfico das novas coordenadas após o escalonamento ótimo dos dados do Quadro 2, obtido junto ao SPSS é apresentado na Figura 1.

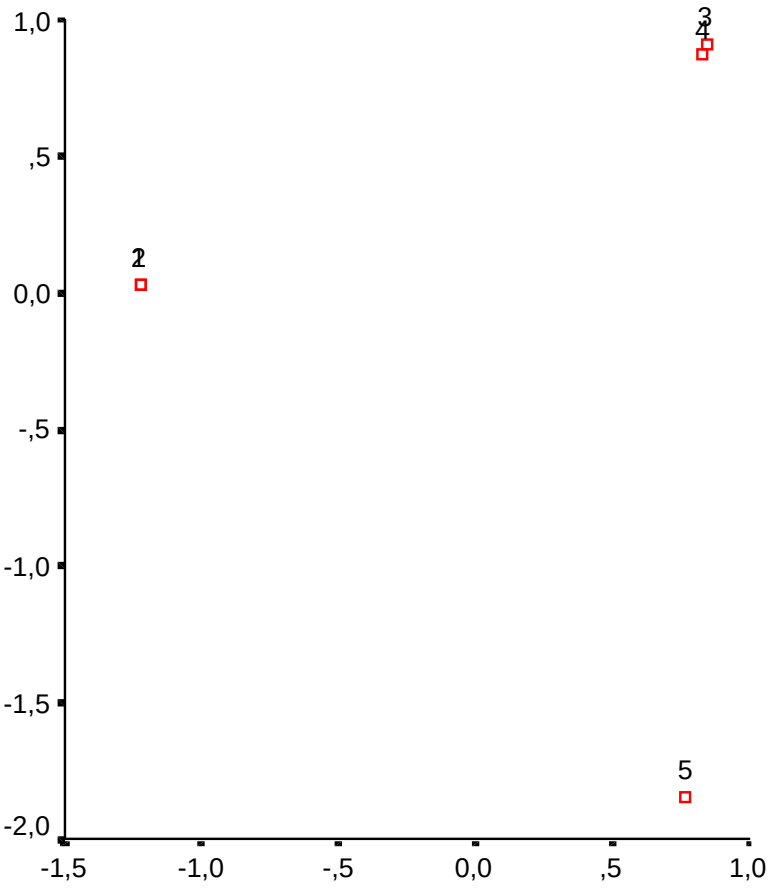

Figura 1 Resultado do escalonamento ótimo

Analisado o gráfico da Figura 1, é possível perceber que os casos 1 e 2 estão bem próximos, assim como os casos 3 e 4, indicando indivíduos que apresentam características semelhantes nas variáveis consideradas. Por exemplo, os indivíduos 1 e 2 são do sexo masculino, da região norte e de cor Parda.

Observação Importante:

Conforme discutido no capítulo 4, quando as variáveis são nominais, a atribuição dos valores para as categorias é arbitrária, não havendo alguma ordinalidade a ser observada. As conclusões da análise serão as mesmas, independente dos valores utilizados.

Para ilustrar, considere o mesmo exemplo de variáveis nominais do Quadro 2, porém atribuindo outros valores, conforme apresentados a seguir:

Variável:	Sexo	Região	Cor
	5 – Masculino	7 – Norte	10 – Preta
	7 – Feminino	6 – Sul	15 – Branca
		4 – Nordeste	20 – Parda
		1 – Sudeste	

Passo 0: Determinar a matriz das observações $\mathbf{X}_{nxm}$

$$\mathbf{X}_{5x3} = \begin{bmatrix} 5 & 7 & 20 \\ 5 & 7 & 20 \\ 7 & 1 & 15 \\ 7 & 4 & 15 \\ 7 & 6 & 10 \end{bmatrix}$$

Passo 1: Criar a matriz Z, padronizando X por coluna

$$\mathbf{Z}_{5x3} = \begin{bmatrix} -1{,}095 & 0{,}784 & 0{,}956 \\ -1{,}095 & 0{,}784 & 0{,}956 \\ 0{,}730 & -1{,}569 & -0{,}239 \\ 0{,}730 & -0{,}392 & -0{,}239 \\ 0{,}730 & -0{,}392 & -1{,}434 \end{bmatrix}$$

Normalizando a matriz **Z**, para $\mathbf{Z}^*$ tem-se que:

$$\mathbf{Z}^{*}_{5x3} = \begin{bmatrix} -0{,}548 & 0{,}392 & 0{,}478 \\ -0{,}548 & 0{,}392 & 0{,}478 \\ 0{,}365 & -0{,}784 & -0{,}120 \\ 0{,}365 & -0{,}196 & -0{,}120 \\ 0{,}365 & -0{,}196 & -0{,}717 \end{bmatrix}$$

Passo 2: Estimativa do Modelo

Calcula-se a matriz de correlações **R** a partir de $\mathbf{Z}^{*}$, através da expressão:

$$\mathbf{R} = \mathbf{Z}^{*t}\mathbf{Z}^{*}$$

$$\mathbf{R}_{3x3} = \begin{bmatrix} 1 & -0{,}716 & -0{,}873 \\ -0{,}716 & 1 & 0{,}352 \\ -0{,}873 & 0{,}352 & 1 \end{bmatrix}$$

Autovalores de **R**, calculados pela equação característica, $\left| \mathbf{R} - \lambda\mathbf{I} \right| = 0$ são:

$\lambda_1 = 2{,}316$

$\lambda_2 = 0{,}657$

$\lambda_3 = 0{,}027$

Os dois primeiros autovalores dão uma variância explicada de $\frac{2{,}316 + 0{,}657}{3} = \frac{2{,}973}{3{,}000} = 99{,}10\%$.

Sendo $\mathbf{D}_k$ a matriz diagonal formada pelos k maiores autovalores, considerando k = 2 componentes, tem-se:

$$\mathbf{D}_2 = \begin{bmatrix} 2{,}316 & 0 \\ 0 & 0{,}657 \end{bmatrix} \quad \text{e sua inversa} \quad \mathbf{D}_2^{-1} = \begin{bmatrix} 0{,}432 & 0 \\ 0 & 1{,}523 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} 0{,}651 & 0{,}074 & 0{,}756 \\ -0{,}505 & 0{,}785 & 0{,}358 \\ -0{,}567 & -0{,}615 & 0{,}549 \end{bmatrix}$$

Sendo **V** a Matriz dos Autovetores associados:

considerando k = 2 componentes , tem-se: $\mathbf{V}_2 = \begin{bmatrix} 0,651 & 0,074 \\ -0,543 & 0,785 \\ -0,567 & -0,615 \end{bmatrix}$

Passo 3: Determinar a Matriz $\mathbf{Z}^{''} = \mathbf{FL}^{t}$

A matriz $\mathbf{F}_{nxk}$, contém os *n* escores relativos aos *k* componentes principais; e a matriz $\mathbf{L}_{mx\,k}$, contém as cargas fatoriais das *m* variáveis em termos dos *k* componentes.

Cálculos auxiliares para determinar $\mathbf{Z}^{''}$

$$\mathbf{L}_j = \sqrt{\lambda_j}\,\mathbf{v}_j$$

$$\mathbf{L} = \begin{bmatrix} 0,990 & 0,060 \\ -0,769 & 0,636 \\ -0,862 & -0,498 \end{bmatrix}$$

$$\mathbf{F} = \mathbf{Z}^{*}\mathbf{L}\mathbf{D}_{\mathrm{K}}^{-1}$$

$$\mathbf{F} = \begin{bmatrix} -0,542 & -0,033 \\ -0,542 & -0,033 \\ 0,461 & -0,636 \\ 0,266 & -0,066 \\ 0,358 & 0,768 \end{bmatrix}$$

$$\mathbf{Z}^{''} = \begin{bmatrix} -0,539 & 0,396 & 0,484 \\ -0,539 & 0,396 & 0,484 \\ 0,418 & -0,759 & -0,081 \\ 0,259 & -0,246 & -0,196 \\ 0,401 & 0,213 & -0,691 \end{bmatrix}$$

Verificando a qualidade do ajuste $\theta^{*} = tr\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)^{t}\left(\mathbf{Z}^{*} - \mathbf{Z}^{''}\right)$

$$\left(\mathbf{Z}^{*} \quad \mathbf{Z}^{''}\right)^{t}\left(\mathbf{Z}^{*} \quad \mathbf{Z}^{''}\right) = \begin{bmatrix} 0,015 & 0,007 & 0,011 \\ 0,007 & 0,003 & 0,005 \\ 0,011 & 0,005 & 0,008 \end{bmatrix}$$

$\theta^* = 0{,}027$

Passo 4: Escalonamento Ótimo

Seja o vetor **B** , contendo os valores de $\mathbf{Z}^*$ ordenados.

Fazendo o escalonamento por variável, tem-se:

Para a variável A:

$$\mathbf{Z}^*_{5x3} = \begin{bmatrix} -0{,}548 & 0{,}392 & 0{,}478 \\ -0{,}548 & 0{,}392 & 0{,}478 \\ 0{,}365 & -0{,}784 & -0{,}120 \\ 0{,}365 & -0{,}196 & -0{,}120 \\ 0{,}365 & 0{,}196 & -0{,}717 \end{bmatrix}$$

$$\mathbf{B}^t = [\,-0{,}548 \quad -0{,}548 \quad 0{,}365 \quad 0{,}365 \quad 0{,}365]$$

Seja o vetor $\hat{\mathbf{Z}}^t$ contendo os elementos de $\mathbf{Z}^{"}$

(Os elementos de $\hat{\mathbf{Z}}^t$ mantém uma ordem em correspondência com a matriz **B**)

$$\mathbf{Z}^{\square\grave{E}t} = [\; -0{,}548 \quad -0{,}548 \quad 0{,}418 \quad 0{,}259 \quad 0{,}401 \;]$$

Criação da Matriz Indicadora : **G**

$$\mathbf{G}^t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Calcula-se agora $\mathbf{Z}^G = \mathbf{G}\left(\mathbf{G}^t\mathbf{G}\right)^{-1}\mathbf{G}^t\,\hat{\mathbf{Z}}^t$

$$\mathbf{Z}^{G} = \begin{bmatrix} -0{,}539 \\ -0{,}539 \\ 0{,}360 \\ 0{,}360 \\ 0{,}360 \end{bmatrix}$$

Recolocam-se os resultados numa nova matriz **X** como foi usado para fazer de **Z** para **B**, e recomeça o processo a partir desta nova matriz **X**.

Recolocando:

$$\mathbf{X} = \begin{bmatrix} -0{,}539 & ..... & ..... \\ -0{,}539 & ..... & ..... \\ 0{,}360 & ..... & ..... \\ 0{,}360 & ..... & ..... \\ 0{,}360 & ..... & ..... \end{bmatrix}$$

O mesmo é feito para as demais variáveis, onde vamos obter que:

$$\mathbf{X} = \begin{bmatrix} -0{,}539 & 0{,}396 & 0{,}484 \\ -0{,}539 & 0{,}396 & 0{,}484 \\ 0{,}360 & -0{,}759 & -0{,}139 \\ 0{,}360 & -0{,}246 & -0{,}139 \\ 0{,}360 & 0{,}213 & -0{,}691 \end{bmatrix}$$

A partir desta nova matriz **X** , repete-se o passo 1 ao 3, até que a convergência do método seja alcançada.

O valor de θ^{*} nesta iteração será:

$\theta^{*} = 0,018$ ⟶ **diminuiu**

E o processo continua até que a diferença de θ entre uma iteração e outra seja desprezível.

O gráfico da Figura 2 é o resultado das novas coordenadas obtidas através do escalonamento ótimo dos dados do Quadro 2, porém utilizando outros valores arbitrários para as categorias.

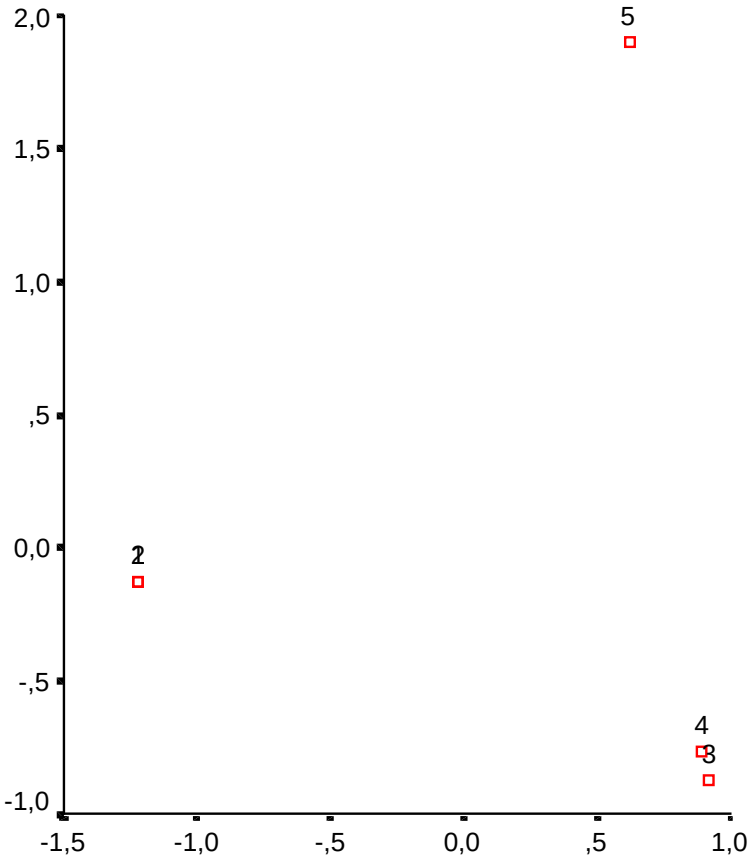

Figura 2 Resultado do escalonamento ótimo

Comparando-se os gráficos das figuras 1 e 2, pode-se observar que eles revelam as mesmas informações em relação aos indivíduos próximos.

Anexo 2 – Resultados da Análise Exploratória de Dados

Resultados para as Variáveis Não Quantitativas

Variável: CapEnx - Capacidade de Enxergar

Tabela 1 - Distribuição de freqüências da variável CapEnx

Cidade	Códigos/Categorias					Total
	1	2	3	4	9	
BNU	14	238	1.458	16.445	62	18.217
CHA	4	167	757	8.565	41	9.534
CRI	9	226	1.143	9.710	33	11.121
ITA	9	209	1.092	8.465	102	9.877
JGS	2	74	559	6.790	20	7.445
JOI	21	509	2.571	25.243	110	28.454
LAG	9	219	1.176	8.204	20	9.628
PAL	4	160	813	5.677	8	6.662
SJO	13	234	1.303	10.042	23	11.615
TUB	5	142	581	5.488	7	6.223
Total	90	2.178	11.453	104.629	426	118.776

O tratamento para o código 9 (*missing)* foi a exclusão da análise para o registro.

Variável: CapOuv - Capacidade de Ouvir

Tabela 2 Distribuição de freqüências da variável CapOuv

Cidade	Códigos/Categorias					Total
	1	2	3	4	9	
BNU	18	88	603	17.440	68	18.217
CHA	10	75	311	9.068	70	9.534
CRI	8	63	418	10.606	26	11.121
ITA	16	63	437	9.285	76	9.877
JGS	5	37	206	7.168	29	7.445
JOI	31	170	1.084	27.057	112	28.454
LAG	10	99	480	9.012	27	9.628
PAL	9	56	223	6.362	12	6.662
SJO	8	83	479	11.016	29	11.615
TUB	4	58	264	5.890	7	6.223
Total	119	792	4.505	112.904	456	118.776

O tratamento para o código 9 (*missing)* foi a exclusão da análise para o registro.

Variável: LerEscrev – Sabe Ler e Escrever

Tabela 3 - Distribuição de freqüências da variável LerEscrev

Cidade	Códigos/Categorias		Total
	1	2	
BNU	17.732	485	18.217
CHA	8.828	706	9.534
CRI	10.623	498	11.121
ITA	9.385	492	9.877
JGS	7.233	212	7.445
JOI	27.472	982	28.454
LAG	8.938	690	9.628
PAL	6.202	460	6.662
SJO	9.873	1.742	11.615
TUB	5.317	906	6.223
Total	111.603	7.173	118.776

Variável: FreqEscola – Freqüenta Escola

Tabela 4 - Distribuição de freqüências da variável FreqEscola

Cidade	Códigos/Categorias				Total
	1	2	3	4	
BNU	1.086	974	15.755	402	18.217
CHA	540	1.096	7.332	566	9.534
CRI	678	496	9.512	435	11.121
ITA	482	434	8.543	418	9.877
JGS	513	448	6.316	168	7.445
JOI	1.804	1.532	24.256	862	28.454
LAG	590	469	8.008	561	9.628
PAL	203	348	5.784	327	6.662
SJO	1.353	2.325	6.769	1.168	11.615
TUB	348	1.551	3.837	487	6.223
Total	7.597	9.673	96.112	5.394	118.776

Variável: EstCivil – Estado Civil

Tabela 5 - Distribuição de freqüências da variável EstCivil

Cidade	Códigos/Categorias						Total
	0	1	2	3	4	5	
BNU	-	9.663	683	499	951	6.421	18.217
CHA	-	5.177	294	190	451	3.422	9.534
CRI	-	6.169	321	297	663	3.671	11.121
ITA	-	4.998	353	316	666	3.544	9.877
JGS	-	4.191	265	135	370	2.484	7.445
JOI	-	16.248	1.020	741	1.505	8.940	28.454
LAG	-	4.891	307	274	639	3.517	9.628
PAL	-	3.251	210	153	336	2.712	6.662
SJO	2.030	3.598	341	278	398	4.970	11.615
TUB	1.021	2.338	145	134	259	2.326	6.223
Total	3.051	60.524	3.939	3.017	6.238	42.007	118.776

Observação: Os casos com código 0 foram excluídos.

Variável: QtsTrabSemana – Quantos trabalhos tinha na semana de 23 a 29 de julho de 2000

Tabela 6 - Distribuição de freqüências da variável QtsTrabSemana

Cidade	Códigos/Categorias			Total
	0	1	2	
BNU	5.739	12.104	374	18.217
CHA	3.141	6.142	251	9.534
CRI	4.365	6.475	281	11.121
ITA	3.931	5.770	176	9.877
JGS	2.257	5.054	134	7.445
JOI	11.631	16.418	405	28.454
LAG	4.251	5.179	198	9.628
PAL	2.568	3.958	136	6.662
SJO	6.516	4.873	226	11.615
TUB	3.462	2.695	66	6.223
Total	47.861	68.668	2.247	118.776

Variável: TrabEra – Como era o trabalho

Tabela 7 - Distribuição de freqüências da variável TrabEra

Cidade	Códigos/Categorias										Total
	0	1	2	3	4	5	6	7	8	9	
BNU	5.739	263	280	7.103	1.530	588	2.543	46	82	43	18.217
CHA	3.141	174	334	2.981	1.055	323	1.362	21	108	35	9.534
CRI	4.365	131	245	3.284	1.303	338	1.326	31	77	21	11.121
ITA	3.931	132	196	2.797	1.054	283	1.405	20	46	13	9.877
JGS	2.257	103	143	3.296	496	183	901	11	32	23	7.445
JOI	11.631	394	556	9.543	2.172	690	3.256	55	108	49	28.454
LAG	4.251	142	272	2.335	1.183	261	1.104	15	49	16	9.628
PAL	2.568	183	167	1.809	840	101	949	9	25	11	6.662
SJO	6.516	136	141	2.435	1.109	220	1.013	18	27	-	11.615
TUB	3.462	68	103	1.215	497	105	650	12	78	33	6.223
Total	47.861	1.726	2.437	36.798	11.239	3.092	14.509	238	632	244	118.776

Variável: ContribInstPrevOf – Era contribuinte de Instituto de Previdência Oficial

Tabela 8 - Distribuição de freqüências da variável ContribInstPrevOf

Cidade	Códigos/Categorias			Total
	0	1	2	
BNU	13.670	2.016	2.531	18.217
CHA	6.829	866	1.839	9.534
CRI	8.175	818	2.128	11.121
ITA	7.256	805	1.816	9.877
JGS	5.836	647	962	7.445
JOI	22.213	2.050	4.191	28.454
LAG	7.253	762	1.613	9.628
PAL	4.899	433	1.330	6.662
SJO	9.574	610	1.431	11.615
TUB	4.953	356	914	6.223
Total	90.658	9.363	18.755	118.776

Variável: HorasTr – Total de Horas Trabalhadas

Tabela 9 - Distribuição de freqüências da variável HorasTr

Cidade	Códigos/Categorias				Total
	0	1	2	3	
BNU	5.739	777	7.891	3.810	18.217
CHA	3.141	567	3.342	2.484	9.534
CRI	4.365	638	3.423	2.695	11.121
ITA	3.931	456	2.978	2.512	9.877
JGS	2.257	289	3.319	1.580	7.445
JOI	11.631	1.294	9.497	6.032	28.454
LAG	4.251	557	2.906	1.914	9.628
PAL	2.568	362	1.931	1.801	6.662
SJO	6.516	470	2.873	1.756	11.615
TUB	3.462	305	1.240	1.216	6.223
Total	47.861	5.715	39.400	25.800	118.776

Variável: Aposent – Era aposentado em Julho de 2000

Tabela 10 - Distribuição de freqüências da variável Aposent

Cidade	Códigos/Categorias			Total
	0	1	2	
BNU	-	2.629	15.588	18.217
CHA	-	1.044	8.490	9.534
CRI	-	1.814	9.307	11.121
ITA	-	1.185	8.692	9.877
JGS	-	923	6.522	7.445
JOI	-	3.641	24.813	28.454
LAG	-	1.371	8.257	9.628
PAL	-	662	6.000	6.662
SJO	2.030	894	8.691	11.615
TUB	1.021	687	4.515	6.223
Total	3.051	14.850	100.875	118.776

Observação: Os casos com código 0 foram excluídos.

Resultados para as Variáveis Quantitativas

Variável: IdadeAnos – Idade (anos completos)

Para analisar variáveis quantitativas será utilizado o histograma. A Figura 3 representa o histograma para a variável idade.

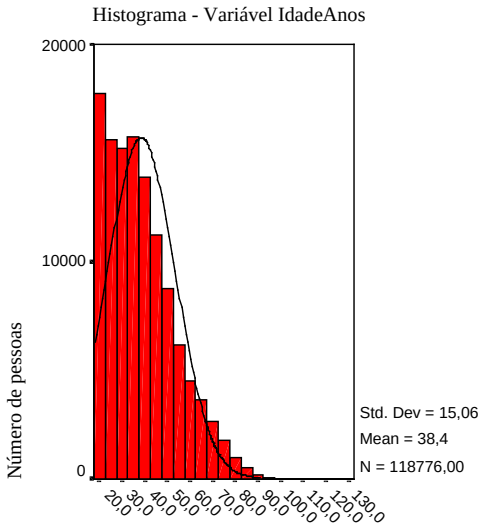

Figura 3 - Histograma da variável Idade

A Tabela 11 apresenta algumas estatísticas da variável Idade. O valor Máximo de 130 anos para idade, provavelmente, foi erroneamente obtido e deve ser eliminado da análise.

Tabela 11 - Medidas resumo para a Variável IdadeAnos

Estatística	Valor
Média	38,40
Mediana	36,00
Moda	18,00
Valor Mínimo	18,00
Valor Máximo	130,00
Desvio Padrão	15,06

Variável: AnosEstudo – Anos de Estudo

O histograma da Figura 4 referente a variável AnosEstudo indica também uma normalidade dos dados obtidos. O casos com códigos 20 e 30, tempo de estudo não determinado e alfabetização de adultos, serão tratados como valores perdidos (*missing values*)

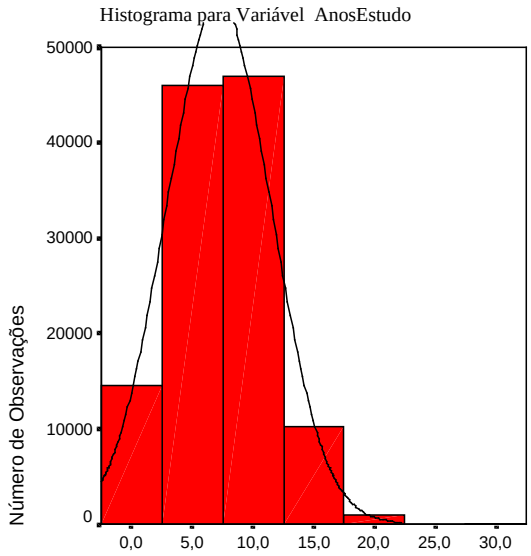

Figura 4 - Histograma para a variável AnosEstudo

Na Tabela 12 são apresentadas algumas estatísticas para esta variável.

Tabela 12 - Medidas resumo para a Variável AnosEstudo

Estatística	Valor
Média	7,18
Mediana	7,00
Moda	4,00
Valor Mínimo	0,00
Valor Máximo	30,00
Desvio Padrão	4,31

Variável: TotRenda – Renda em Reais

O histograma da Figura 5 é referente a variável TotRenda. Pode-se observar uma concentração de observações no início do gráfico e valores extremos muito altos no eixo horizontal, podendo ser valores discrepantes.

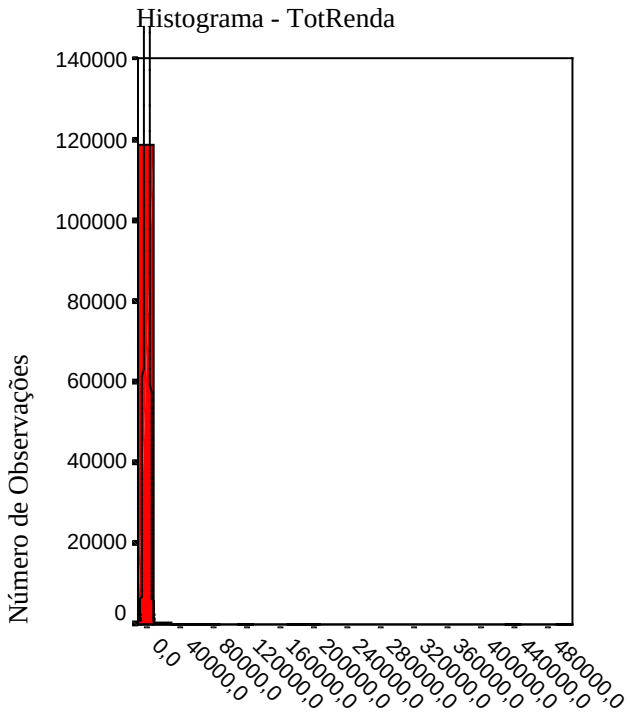

Figura 5 - Histograma para a variável Idade

A Tabela 13 confirma a suspeita de valores discrepantes através do Valor Máximo de 500.350,00.

Tabela 13 - Medidas resumo para a Variável TotRenda

Estatística	Valor
Média	559,74
Mediana	300,00
Moda	0,00
Valor Mínimo	0,00
Valor Máximo	500.350,00
Desvio Padrão	2.430,07

Além deste valor foram encontrados outros valores, considerados discrepantes:

Registro - Valor

98.866 – 500.350,00

35.751 – 430.000,00

4.102 - 200.000,00

Estes valores foram excluídos da análise, assim como os salários acima de R$ 20.000,00.

Após a exclusão foi reconstruído o histograma para esta variável e apresentado na Figura 6

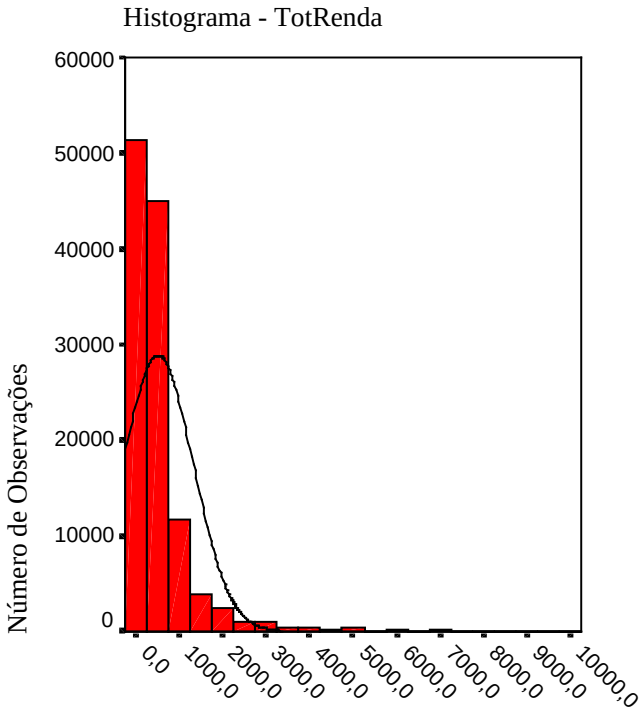

Figura 6 - Histograma para a variável Idade

Anexo 3 – Escalonamento ótimo para as variáveis qualitativas

Tabela 14 Valores do escalonamento ótimo para a variável CapEnx

Valores originais	Valores do EO
1 - incapaz	-7,678
2- grande dificuldade permanente	-5,008
3 - alguma dificuldade permanente	-2,338
4 - nenhuma dificuldade permanente	0,332

Tabela 15 Valores do escalonamento ótimo para a variável CapOuv

Valores originais	Valores do EO
1 - incapaz	-10,452
2- grande dificuldade permanente	-8,521
3 - alguma dificuldade permanente	-4,166
4 - nenhuma dificuldade permanente	0,188

Tabela 16 Valores do escalonamento ótimo para a variável QtTrabSemana

Valores originais	Valores do EO
0 - Não	-1,264
1 - Um	0,646
2 - Dois ou Mais	2,556

Tabela 17 Valores do escalonamento ótimo para a variável HorasTr

Valores originais	Valores do EO
0 - Não trabalhou	-1,213
1 - até 30 horas	-0,384
2 - acima de 30 e menos e até 44 horas	0,445
3 - acima de 44 horas	1,273

Tabela 18 Valores do escalonamento ótimo para a variável LerEscrev

Valores originais	Valores do EO
1 - sim	-0,139
2 - não	7,187

Tabela 19 Valores do escalonamento ótimo para a variável FreqEscola

Valores originais	Valores do EO
1 - sim, rede particular	-3,203
2 - sim, rede pública	-1,412
3 - não, já freqüentou	0,378
4 - nunca freqüentou	0,041

Tabela 20 Valores do escalonamento ótimo para a variável EstCivil

Valores originais	Valores do EO
1 - Casado judicialmente	-0,918
2 - desquitado (a) ou separado (a)	-0,391
3 - divorciado (a)	0,136
4 - viúvo (a)	0,663
5 - solteiro (a)	1,190

Tabela 21 Valores do escalonamento ótimo para a variável TrabEra

Valores originais	Valores do EO
0 - não se aplica	-1,142
1 - trabalhador doméstico c/ carteira assinada	-0,680
2 - trabalhador doméstico s/ carteira assinada	-0,218
3 - empregadp c/ carteira assinada	0,245
4 - empregado s/ carteira assinada	0,707
5 - empregador	1,169
6 - conta-própria	1,631
7 - aprendiz ou estagiário sem remuneração	2,094
8 - não remunerado em ajuda a membro do domicílio	2,556
9 - trabalhador na produção para o próprio consumo	3,018

Tabela 22 Valores do escalonamento ótimo para a variável ContribInstPrevOf

Valores originais	Valores do EO
0 - não se aplica	-0,554
1 - sim	0,772
2 - não	2,097

Tabela 23 Valores do escalonamento ótimo para a variável Aposent

Valores originais	Valores do EO
1 - sim	-2,795
2 - não	0,358